



XI BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS

Information Systems:
A Computer Socio-Technical Perspective

May 26-29, 2015
Goiânia - Goiás - Brazil



**Proceedings of the
XI Brazilian Symposium on Information Systems**

**Information Systems:
A Computer Socio-Technical Perspective**

MAY 26-29, 2015
GOIÂNIA - GOIÁS - BRAZIL

Vol. 1

Published by:

Instituto de Informática

Universidade Federal de Goiás

<http://www.inf.ufg.br/sbsi2015>

Credits:

Cover Design: Ascom/UFG

Proceedings Layout: Sérgio Teixeira de Carvalho, Sean Wolfgang Matsui Siqueira, Leandro Luís Galdino de Oliveira

Dados Internacionais de Catalogação na Publicação (CIP)
GPT/BC/UFG

S612a Simpósio Brasileiro de Sistemas de Informação (20. : 2015 :
Goiânia, GO)

Anais [do] XX Simpósio Brasileiro de Sistemas de Informação
[recurso eletrônico] / coordenação do Comitê de Programa: Sean
Wolfgang Matsui Siqueira, Sérgio Teixeira de Carvalho ; coordena-
ção geral: Vinícius Sebba Patto, Valdemar Vicente Graciano Neto ;
realização: UFG/Instituto de Informática ; promoção: Sociedade
Brasileira de Computação. – Goiânia : UFG, Instituto de Informáti-
ca, 2015.

Tema central: Sistemas de informação: uma visão sociotécnica
da computação.

Conteúdo: v.1. Trilhas técnicas. – v.2. Workshop de iniciação
científica em sistemas de informação (II WICSI). – v.3. Workshop
de teses e dissertações em sistemas de informação (VIII WTDSI). –
v.4. Minicursos.

Disponível em: <<http://www.inf.ufg.br/sbsi2015/anais/>>

1. Sistemas de recuperação da informação – Congressos. 2. Tecnologia
– Serviços de informação – Congressos. 3. Internet na
administração pública – Congressos. I. Siqueira, Sean Wolfgang
Matsui . II. Carvalho, Sérgio Teixeira de. III. Patto, Vinícius Sebba.
IV. Graciano Neto, Valdemar Vicente. V. Universidade Federal de
Goiás. Instituto de Informática. VI. Sociedade Brasileira de Com-
putação. VII. Título. VIII. Título: Sistemas de informação: uma
visão sociotécnica da computação.

CDU: 004.65

XI SBSI

XI Brazilian Symposium on Information System
Information Systems: A Computer Socio-Technical Perspective
May 26-29, 2015
Goiânia, Goiás, Brazil.

Committees

General Organization:

Vinícius Sebba Pato (UFG)
Valdemar Vicente Graciano Neto (UFG)

Chairs of the SBSI 2015 Program Committee:

Sean Wolfgang Matsui Siqueira (UNIRIO)
Sérgio Teixeira de Carvalho (UFG)

Steering Committee:

Adriana Clericuzi (UFPB)
Bruno Zarpelão (UEL)
Célia Ghedini Ralha (UnB)
Fernanda Baião (UNIRIO)
José Maria David (UFJF)
Renata Araujo (UNIRIO)
Rita Suzana Maciel (UFBA)

Chairs of the Track “Experience Reports in Industry and Case Studies” (RESI):

Cláudia Cappelli (UNIRIO)
Arnaldo Alves Ferreira Junior (UFG)

Chairs of the Track “IS Applications and Tools” (ASI):

Sean Wolfgang Matsui Siqueira (UNIRIO)
Marcel Ferrante Silva (UFG)

Chairs of the Track “Information Systems Education” (ESI):

Fátima de L. Santos Nunes (EACH-USP)
Vinícius Sebba Pato (UFG)

Chairs of the Track “Government Information Systems” (SIG):

Cláudia Cappelli (UNIRIO)
Renata Araújo (UNIRIO)
Dalton Lopes Martins (UFG)

Chairs of the Track “Business Process Management” (BPM):

Flávia Maria Santoro (UNIRIO)
Fernanda Baião (UNIRIO) Juliano Lopes de Oliveira (UFG)

SBSI 2015 Program Committee:

Adolfo Duran (UFBA)
Adriana Cesário de F. (UNIRIO)
Alexandre Cidral (Univille)

Alexandre Correa (UNIRIO)
Ana Paula Ambrósio (UFG)
Andrea Magdaleno (UNIRIO)
Andreia Malucelli (PUCPR)
Andrew Costa (PUC-Rio)
Antonio Tadeu Azevedo Gomes (LNCC)
Astério Tanaka (UNIRIO)
Avanilde Kemczinski (UDESC)
Carina Alves (UFPE)
Carine Webber (UCS)
Carla Merkle Westphall (UNIRIO)
Carlos Alberto Vieira Campos (UNIRIO)
Carlos Eduardo Santos Pires (UFCG)
Carlos Kamienski (UFABC)
Carlos Santos Jr. (UnB)
Celia Ralha (UnB)
Claudia Cappelli (UNIRIO)
Clodis Boscarioli (UNIOESTE)
Clodoaldo Lima (USP)
Cristiano Maciel (UFMT)
Dalton Martins (UFG)
Daniela Barreiro Claro (UFBA)
Debora Muchaluat-Saade (UFF)
Debora Paiva (UFMS)
Denis Silveira (UFPE)
Edmir Prado (USP)
Edmundo Spoto (UFG)
Edson Cáceres (UFMS)
Eduardo Almentero (UFRRJ)
Ellen Francine Barbosa (USP)
Elvis Fusco (UNIVEM)
Eraclito Argolo (UFMA)
Fábio Bezerra (UFRA)
Fatima Nunes (USP)
Fernanda Baião (UNIRIO)
Fernanda Campos (UFJF)
Fernanda Lima (UnB)
Fernando Braz (IFC)
Flavia Santoro (UNIRIO)
Flávio Soares Corrêa da Silva (USP)
Geiza M. Hamazaki da Silva (UNIRIO)
Glauro Carneiro (UNIFACS)
Gleison Santos (UNIRIO)
Ildeberto Rodello (USP)
Jairo Souza (UFJF)
João Porto de Albuquerque (USP)
Jonice Oliveira (UNIRIO)
Jorge Barbosa (Unisinos)
José David (UFJF)
José Maldonado (USP)
José Viterbo (UFF)
Josivaldo Araujo (UFPA)
Jugurta Lisboa Filho (UFV)
Juliano Lopes de Oliveira (UFG)
Kate Revoredo (UNIRIO)
Leandro Augusto Silva (Mackenzie)
Leonardo Azevedo (UNIRIO)
Lisandro Zambenedetti Granville (UFRGS)
Luciana Martimiano (UEM)

Luciano Digiampietri (USP)
Lucineia Heloisa Thom (UFRGS)
Marcelo Fantinato (USP)
Marcelo Ladeira (UnB)
Marcelo Morandini (USP)
Marcelo Yamaguti (PUCRS)
Marcos Chaim (USP)
Maria Claudia Cavalcanti (IME-RJ)
Maria Istela Cagnin (UFMS)
Mariano Pimentel (UNIRIO)
Melise Paula (UNIFEI)
Morganna Diniz (UNIRIO)
Olinto Furtado (UFSC)
Patricia Vilain (UFSC)
Paulo Pires (UNIRIO)
Plínio Leitão Júnior (UFG)
Regina Braga (UFJF)
Rejane Frozza (UNISC)
Renata Araujo (UNIRIO)
Renata S.S. Guizzardi (UFES)
Ricardo Choren (IME-RJ)
Ricardo Falbo (UFES)
Ricardo Machado (Universidade do Minho)
Ricardo Rabelo (UFSC)
Rita Suzana Pitangueira Maciel (UFBA)
Roberto Pereira (UNICAMP)
Rodolfo Barros (UEL)
Rodolfo Resende (UFMG)
Rogério de Carvalho (IFF)
Rosângela Penteado (UFSCAR)
Sarajane Peres (USP)
Sean Siqueira (UNIRIO)
Sidney Lucena (UNIRIO)
Sílvio Cazella (Unisinos)
Simone Martins (UFF)
Tania Tait (UEM)
Thais Vasconcelos Batista (UFRN)
Valter Camargo (UFSCAR)
Vanessa Nunes (UnB)
Vânia Maria Dias (UNIRIO)
Vaninha Vieira (UFBA)
–Wagner Meira Jr. (UFMG)
Wallace Ugulino (PUC-Rio)

Additional Reviewers:

Adailton Araújo (UFG)
Adler Diniz de Souza (UNIFEI)
Adriana Silveira de Souza (UFG/PUC-GO/Estratégia TI)
Alessandro Machado de Araújo (UFG)
Alexandre Bragança (ISEP)
Aline Paes (UFF)
Anderson Marcolino (USP)
Augusto Baffa (PUC-Rio)
Auri Marcelo Rizzo Vincenzi (UFG)
Bernardo Pereira Nunes (PUC-Rio)
Cassio Rodrigues (UFG)
Cleber Correa (USP)
Cristiane Ramos (UnB)
Cristina Cerdeiral (UFRJ)

Davi Viana (UFAM)
Deller James (UFG)
Dimas C. Nascimento (UFCG)
Edson Moreira (USP)
Eduardo Albuquerque (UFG)
Eduardo Motta (PUC-Rio)
Eunice Nunes (UFMT)
Fabiana Mendes (UnB)
Fernando Castilho (PUCRS)
Fernando Pinhati Júnior (UNIRIO)
Flavio Horita (USP)
Francisco Vasconcellos (UFMT)
Hebert da Silva (UFG)
Helton Biscaro (USP)
Jivago Medeiros (UFMT)
Leila Bergamasco (USP)
Livia Castro Degrossi (USP)
Luis Alves (IPB)
Luiz Fernando Assis (USP)
Luiza de Paula (UNRIO)
Maiara Cancian (UFSC)
Marcelo Camanho (PUC-Rio)
Marcelo Quinta (UFG)
Maria Helena Braz (Universidade de Lisboa - ICIST)
Natália Chaves Lessa Schots (UFRJ)
Nemésio Freitas Duarte Filho (USP)
Patricia Mendonça (UFMT)
Rafael Oliveira (USP)
Raul Eduardo Simoni Castanhari (USP)
Ricardo da Rocha (UFG)
Roberto Rocha (USP)
Sofia Costa (USP)
Suelane Garcia (USP)
Taciana Kudo (UFG)
Tadeu Classe (UFJF)
Thiago Procaci (UNIRIO)
Thierson Rosa (UFG)
Valdemar Neto (UFG)

Organization

INF/UFG – Instituto de Informática/ Universidade Federal de Goiás

Promotion

SBC – Sociedade Brasileira de Computação

Support

SECTEC-GO – Secretaria de Ciência, Tecnologia e Inovação do Estado de Goiás

Sponsors

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

FAPEG – Fundação de Amparo à Pesquisa do Estado de Goiás

Faculdades ALFA

Global RH Solutions

SBSI 2015 is an AIS affiliated conference, in cooperation with ACM SIGAPP and ACM SIGMIS

AIS – Association for Information Systems

ACM SIGAPP – Association for Computing Machinery Special Interest Group on Applied Computing

ACM SIGMIS – Association for Computing Machinery Special Interest Group on Management Information Systems

Preface

The Brazilian Symposium on Information Systems (SBSI) provides a meeting point for researchers, practitioners and students in the area of Information Systems (IS). The main objective of the symposium is to share scientific knowledge and interests and to establish strong professional and academic ties among the participants. The objective is to promote discussions on the Information Systems (IS) area considering its whole multidisciplinary approach and applications in the main different knowledge areas, welcoming a wide range of topics related to IS.

This is the 11th edition of the Brazilian Symposium on Information Systems – SBSI 2015. The symposium theme typically changes from year to year. In 2015, the symposium theme is: Information Systems: a computer socio-technical perspective. In special, this edition focuses on the advances of the IS community aimed at integrating the diverse computer techniques and technologies to the way the society lives, works and communicates. This socio-technical approach is a broader view of IS, involving not only technology and organizations, but also people and society as a whole. This research field has the potential to develop and improve the services of a country or region, through the conception of socio-technical systems (such as health systems, security systems, education systems, and energy and transportation systems, among others).

The SBSI is a traditional and important scientific meeting of the Brazilian Computer Society (SBC), and it is growing in importance. It started in 2004, and in the first three years (from 2004 to 2006), the SBSI was held in different states of the south of the country. Then, in 2008, it was located at the southeast and after that each year it has been held in a different region in Brazil. In 2015 we searched for improving the SBSI's international visibility, then SBSI 2015 is in cooperation with ACM SIGAPP and SIGMIS, and it is also an AIS affiliated conference.

The Brazilian Computer Society (SBC) is a “scientific association, without profit organization, that brings together researchers, teachers, students and professionals engaged in scientific research, education and technological development in the general area of computing”. More information on <http://www.sbc.org.br>.

The SBC Special Committees (ECs) “gather associated members with common interests in a particular subfield of Computing. Each Special Committee also organizes their respective symposium and regularly holds its general assembly”. The Special Committee on Information Systems (CESI) is “the commission responsible for representing the Information Systems community in the Brazilian Computer Society (SBC). The Brazilian Symposium on Information Systems (SBSI) is the main national event of CESI”.

ACM stands for Association for Computing Machinery, “the world’s largest educational and scientific computing society, which delivers resources that advance computing as a science and a profession. ACM provides the computing field’s premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources”. The SBSI 2015 proceedings are going to be available at ACM Digital Library.

The SBSI 2015 was approved for in-cooperation with two ACM's Special Interest Groups: SIGAPP and SIGMIS.

The SIGAPP - ACM Special Interest Group on Applied Computing - is “the ACM's primary applications-oriented SIG. It offers practitioners and researchers the opportunity to share mutual interests in innovative applications, technology transfer, experimental computing, strategic research, and the management of computing. This SIG also promotes widespread cooperation among business, government, and academic computing programs”. More information on <http://www.sigapp.org>.

The SIGMIS - ACM Special Interest Group on Management Information Systems – “promotes best-practice and research in the management of information systems and technologies and the use of these systems and technologies. SIGMIS is a sponsor of several conferences on information systems and technology. As one of the oldest of ACM's SIGS, SIGMIS traces its beginnings back to 1961, and for decades has been instrumental in defining and developing the field of management and information systems”. More information on <http://sigmis.org>.

The Association for Information Systems (AIS) “serves society through the advancement of knowledge and the promotion of excellence in the practice and study of information systems. AIS is the premier professional association for individuals and organizations who lead the research, teaching, practice, and study of information systems worldwide”. More information on <http://aisnet.org>.

SBSI 2015 was held in Goiânia, from May 26 to 29th, 2015. It was organized by the Institute of Informatics of the Federal University of Goiás. The SBSI 2015 schedule included many activities with high technical-scientific level, covering different aspects, including:

- Technical sessions for the presentation of selected scientific articles;
- Invited talks, with national and international experts;
- Panel with representatives of academia, industry and government;
- Workshop of Theses and Dissertations in Information Systems (WTDSI);
- Workshop on Undergraduate Research in Information Systems (WICSI);
- Selected short courses.

The technical sessions were organized in different tracks:

- MAIN TRACK

Chairs: Prof. Sean W. M. Siqueira (UNIRIO) and Prof. Sérgio T. Carvalho (UFG)

- TRACK 1 - “Experience Reports in Industry and Case Studies”

Chairs: Prof. Claudia Cappelli (UNIRIO) and Prof. Arnaldo Alves Ferreira Junior (UFG)

- TRACK 2 - “IS Applications and Tools”

Chairs: Prof. Sean Siqueira (UNIRIO) and Prof. Marcel Ferrante Silva (UFG)

- TRACK 3 - “Information Systems Education”

Chairs: Prof. Fátima de Lourdes dos Santos Nunes Marques (USP) and Prof. Eliany Alvarenga de Araújo (UFG)

- TRACK 4 - “Government Information Systems”

Chairs: Prof. Claudia Cappelli (UNIRIO), Prof. Renata Araujo (UNIRIO) and Prof. Dalton Lopes Martins (UFG)

- TRACK 5 - “Business Process Management”

Chairs: Prof. Flávia Maria Santoro (UNIRIO), Prof. Fernanda Baião (UNIRIO) and Prof. Juliano Lopes de Oliveira (UFG)

The Proceedings of the SBSI 2015 have excellent papers that discuss and develop innovative themes exploring knowledge of various areas related to Information Systems. There were 313 submissions of regular papers written in English and Portuguese (main track: 168 papers; track 1: 36 papers; track 2: 42 papers; track 3: 16 papers; track 4: 25 papers; and track 5: 26 papers), of which 302 were validated (main track: 166 papers; track 1: 32 papers; track 2: 40 papers; track 3: 15 papers; track 4: 24 papers; and track 5: 25 papers). From this total, 101 papers were accepted (91 full papers and 10 short papers) (main track: 50 papers; track 1: 12 papers; track 2: 12 papers; track 3: 7 papers; track 4: 13 papers; and track 5: 7 papers), which represents 33,44% of acceptance rate. Full papers contain between 8 (eight) and 10 (ten) pages, while short papers contain from four (4) and five (5) pages.

We appreciate the excellent cooperation of the track chairs of the SBSI 2015 that with their commitment and dedication allowed the organization of this symposium. Along with the track chairs, we thank the members of the Program Committee and the additional reviewers, which allowed a rigorous evaluation process, in which all the submitted papers were evaluated by researchers with recognized experience in the field. In addition, all revisions were analyzed by the chairs to the final decision.

We also thank the members of the Special Committee on Information Systems (CESI), also the SBSI 2015 Steering Committee that much contributed to every stage of symposium organization: Célia Ralha (UnB) – coordinator, José Maria David (UFJF) – vice-coordinator, Rita Suzana (UFBA), Adriana Clericuzi (UFPB), Renata Araujo (UNIRIO), Fernanda Baião (UNIRIO) and Bruno Zarpelão (UEL).

This symposium would not have been possible without the efforts of several people involved in its organization. Thus, we thank all the organizing committee of the SBSI 2015, including the wonderful staff of the UFG, and to all the people who contributed to the success of this event. Special thanks to all the authors for choosing our symposium as a forum for submission and publication of their work. We also thank everyone who supported us in coordinating the sessions and all participants and sponsors who promote this community.

Goiânia, May 2015.

Sean W. M. Siqueira (UNIRIO) and Sérgio T. Carvalho (UFG)

Chairs of the SBSI 2015 Program Committee

Short biography of the Chairs of the Program Committee



Sean Wolfgang Matsui Siqueira is a professor at the Department of Applied Informatics, Federal University of the State of Rio de Janeiro (UNIRIO), Brazil, where he teaches courses in Information Systems, Databases, Web Science and Social & Semantic Web. He holds a M.Sc. (1999) and a Ph.D. (2005) in Computer Science, both from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil. His research interests include knowledge representation and management; collaborative systems; social web; semantic web; ontologies; information integration; semantic models; user models; and learning & education. He has experience in the Computer Science area, with focus on Information Systems and Technology Enhanced Learning. He has participated in some international research projects and has written more than 130 papers for conferences, journals, and book chapters. He was the coordinator of the Graduate Program in Information Systems at UNIRIO from July 2012 to September 2014, and nowadays is the editor-in-chief of the iSYS: Brazilian Journal of Information Systems.



Sérgio Teixeira de Carvalho is a professor at the Institute of Informatics, Federal University of Goiás (UFG), Brazil, since 2006, where teaches courses in Computer Science, Information Systems and Software Engineering. He has taught in various colleges and university centres in Computer Science area. For several years he served mainly as software project coordinator, and held positions such as Director of the Information Systems, responsible for the development and maintenance of systems, and Director of Technology Support, responsible for hardware and software infrastructure. For over sixteen years, he served as IT professional on various public and private companies, exercising activities on software development, database management and network and systems management. He holds a BSc. in Computer Science from the UFG, a MSc. and a DSc. in Computer Science, both from the Fluminense Federal University (UFF). He has experience in Distributed Systems and Software Engineering areas, and his main fields of activity are ubiquitous computing, with focus on health care applications, software architecture, dynamic software product lines, and self-adaptive architectures.

TECHNICAL TRACKS

Best Paper Candidates

1

- 1 Capacitated vehicle routing system applying Monte Carlo Methods - MAIN TRACK
Rômulo Augusto de Carvalho Oliveira (USP), Karina Valdivia Delgado (USP)
- 9 Recognition of Compromised Accounts on Twitter - MAIN TRACK
Rodrigo Augusto Igawa (UEL), Alex Marino Gonçalves de Almeida (UEL), Bruno Bogaz Zarpelão (UEL), Sylvio Barbon Jr (UEL)
- 15 Evaluation of a Maturity Model for Agile Governance in ICT using Focus Group - MAIN TRACK
Humberto Rocha de Almeida Neto (UFPE), Edviges Mariza Campos de Magalhães (DATAPREV), Hermano Perrelli de Moura (UFPE), José Gilson de Almeida Teixeira Filho (UFPE), Claudia Cappelli (UNIRIO), Luiz M. Fraga Martins (CISUC)
- 23 Applying a Variation of the Ant Colony Optimization Algorithm to Solve the Multiple Traveling Salesmen Problem to Route the Teams of the Electric Power Distribution Companies - MAIN TRACK
Denilson F. Barbosa (UTFPR), Carlos N. Silla Jr. (UTFPR), André Y. Kashiwabara (UTFPR)
- 31 Learning Difficulties on Computing Courses: Cognitive Processes Assessment Methods Research and Application - TRACK "INFORMATION SYSTEMS EDUCATION"
Vinicius Vieira Pessoni (UFG), Fernando Marques Federson (UFG), Auri Marcelo Rizzo Vincenzi (UFG)

Main Track

39

MT1 - Information Systems Tools

39

- 39 A Tool for Searching in Unstructured Code AspectJ
Rodrigo Castro Gil (UFESM), Eduardo Kessler Piveta (UFESM), Deise de Brum Saccol (UFESM), Cristiano De Faveri (UFESM)
- 47 Automatic Tuning of GRASP with Path-Relinking in data clustering with F-Race and iterated F-Race
Júlio Cesar da Silva (UNIFEI), Rafael de M. D. Frinhani (UNIFEI), Ricardo Martins A. Silva (UFLA), Geraldo Robson Mateus (UFMG)
- 55 GiveMe Views: A Support Tool to Software Evolution Based on Historical Data Analysis
Jacimar F. Tavares (UFJF), José Maria N. David (UFJF), Marco Antônio P. Araújo (UFJF), Regina Braga (UFJF), Fernanda Campos (UFJF), Glaucio Carneiro (UNIFACS)
- 63 Android Augmented Reality Applications in Extensible, Flexible, and Adaptable Architecture
Tiago Araújo (UFPA), Nikolas Carneiro (UFPA), Brunelli Miranda (UFPA), Carlos Santos (UFPA), Bianchi Meiguins (UFPA)

MT2 - Solutions for Systems-of-Systems and Software Ecosystems

71

- 71 Waldo: Data Producers Registry and Discovery Service for Smart Cities Middleware
Marcelo Iury S. Oliveira (UFRPE), Kiev Santos da Gama (UFPE), Bernadette Farias Lóscio (UFPE)
- 79 Reference Software Architecture for Government Information Systems
Maurício Serrano (UnB), Milene Serrano (UnB), André Cruz Cavalcante (UnB)
- 83 Software Startup Ecosystems: Initial Results in the state of Pará
Nágila Natália de Jesus Torres (UFPA), Cleidson R. B. de Souza (UFPA, ITV)
- 87 A Reference Architecture for a Crowdsensing Platform in Smart Cities
Herbert B. M. Diniz (UFPE), Emanuel C. G. F. Silva (UFPE), Kiev Santos da Gama (UFPE)

MT3 - Prediction Methods and Recommender Systems

95

- 95 Prediction of Artists Rankings by Regression
Felipe L. M. Faria (UFOP), Álvaro R. Pereira Jr. (UFOP), Luiz H. C. Merschmann (UFOP)
- 103 Multimedia Content Recommendation in Digital Convergence Environments: An Approach Based on Data Mining and Semantic Web
Priscilla Kelly M. Vieira (UFPB), Natasha Queiroz Lino (UFPB)

- 111 A File Recommendation Model For Cloud Storage Systems
Ricardo Batista Rodrigues (UFPE), Carlo M.R. da Silva (UFPE), Frederico A. Durão (UFBA), Rodrigo E. Assad (UFRPE), Vinicius C. Garcia (UFPE), Silvio L. Meira (UFPE)
- 119 Prediction System of Bus Arrival Time Based on Historical Data Using Regression Models
Kássio R. Coquita (UFPE), Arley R. R. Ristar (UFPE), Adriano L. I. Oliveira (UFPE), Patrícia C. A. R. Tedesco (UFPE)
- 127 Evaluation of Consumer Shopping Behavior in M-Commerce and E-Commerce
Felipe Lemos (PUC-Minas), Luís Fabrício Góes (PUC-Minas)
- 135 Recognition of trees using Computer Vision to prevent interruptions in the Power Distribution System
Heuber G. F. de Lima (UFG), (UFG), Anderson da S. Soares (UFG), Gustavo T. Laureano (UFG)
- 139 A Link Prediction System in Social Networks
Luciano Antonio Digiampietri (USP), William Takahiro Maruyama (USP), Caio Rafael do Nascimento Santiago (USP), Jamison José da Silva Lima (USP)
- 147 Assesment of Features Influencing the Voting for Opinions' helpfulness about Services in Portuguese
Augusto C. S. Martins (UTFPR), Cesar A. Tacla (UTFPR)

MT4 - Database, Information Retrieval, Data Mining, and Knowledge Discovery

155

- 155 Data Mining Techniques Assessment for Shutdowns Prediction of Electric Power Systems
Anderson Trindade Maia (UFPA), Jefferson Magalhães Moraes (UFPA), Yomara Pinheiro Pires (UFPA), Aldebaro Barreto da Rocha Klautau Júnior (UFPA), Daniel Martins (UFPA)
- 163 An Unsupervised Method for Ontology Population from Textual Sources on the Web
Fábio Lima (UFBA), Hilário Oliveira (UFPE), Laís Salvador (UFBA)
- 171 Design of the Outer-tuning Framework: Self-tuning and Ontology for Relational Databases
Rafael Pereira de Oliveira (PUC-Rio), Sergio Lifschitz (PUC-Rio), Ana Carolina Almeida (PUC-Rio), Edward Hermann Haeusler (PUC-Rio)
- 179 Application of Knowledge Discovery in Databases in Evapotranspiration Estimation: an Experiment in the State of Rio de Janeiro
Fernando Xavier (UNIRIO), Asterio Kiyoshi Tanaka (UNIRIO), Kate Cerqueira Revoredo (UNIRIO)
- 187 Methodology for Generation of Business Intelligence Platforms to University Evaluation Committees
Caio Eduardo Ribeiro (PUC-Minas), Anna Izabel J. Tostes Ribeiro (PUC-Minas)
- 195 Quantitative Temporal Association Rule Mining by Genetic Algorithm
Sérgio F. da Silva (UFG), Marcos A. Batista (UFG), Agma J. M. Traina (USP)

MT5 - Artificial Intelligence for Information Systems

203

- 203 Machine Learning in Textual Content-Based Recommendation Systems: A Systematic Review
Lucas F. Brunialti (USP), Sarajane M. Peres (USP), Valdinei Freire (USP), Clodoaldo A. M. Lima (USP)
- 211 Design and Implementation of an Artificial Neural Network Applied to Finger Bad-positioning Detection on Touchless Multiview Fingerprints Devices
Cauê Zaghetto (UnB), Luiz Henrique M. Aguiar (UnB), Alexandre Zaghetto (UnB)

MT6 - Evidence-based Studies

219

- 219 A Comparative Study of Feature Level Fusion Strategies for Multimodal Biometric Systems Based on Face and Iris
Daniel M. M. da Costa (USP), Henrique Passos (USP), Sarajane Marques Peres (USP), Clodoaldo A. M. Lima (USP)
- 227 Influences of Organizational Culture in the Adoption of Agile Methodologies in Information Systems Development - A Systematic Mapping
Gerson Correia da Silva Junior (UFG), João Paulo Amaral (UFG), Patrícia Gomes Fernandes Mat-subara (UFG), Valdemar Vicente Graciano Neto (UFG)
- 235 Maturing, Consolidation and Performance of NoSQL Databases - Comparative Study
Vanessa Cristina Oliveira de Souza (UNIFEI), Marcus Vinícius Carli dos Santos (UNIFEI)
- 243 Ontology Extraction from Stories: An Exploratory Study in Storytelling
Valdemar T. F. Confort (UNIRIO), Kate Revoredo (UNIRIO), Fernanda Araujo Baião (UNIRIO), Flávia Maria Santoro (UNIRIO)
- 251 Indoor location systems in emergency scenarios - A Survey
Arivan S. Bastos (UFBA), Vaninha Vieira (UFBA), Antonio L. Apolinário Jr. (UFBA)
- 259 Software Processes Improvement in light of Cognitive Biases: A Cross-Case Analysis
José Adson O. G. da Cunha (UFPE), José Jorge L. Dias Jr. (UFPB), Livia Maria R. de V. Cunha (DATAPREV), Hermano Perrelli de Moura (UFPE)

MT7 - Management, Governance, and Government **267**

- 267 Data Governance in Brazilian Organizations
André Montoia Barata (USP), Edmir Parada Vasques Prado (USP)
- 275 Analysis of Scientific Collaboration Network as a Management Tool for Graduate Programs
Aurélio Ribeiro Costa (UnB), Célio Ghedini Ralha (UnB)
- 283 A Context-sensitive Metacognitive Conflict Management Approach for Collaborative Systems
Heremita Brasileiro Lira (IFPB), Patricia Azevedo Tedesco (UFPE)
- 291 Automated Support for Scrum Projects Sprint Planning
Hanna Tátia de Sousa (IFES), Thanner S. Silva (IFES), Paulo S. Santos Jr. (IFES), Rodrigo F. Calhau (IFES), Mateus Barcellos Costa (IFES)

MT8 - Security in Information Systems **299**

- 299 Proposal for Simplified Security Model for Small and Medium Business
Gonçalo M. da Silva Neto (UFRPE), Gliner Dias Alencar (UFPE / IBGE), Anderson Apolonio L. Queiroz (IFRN)
- 307 Analysis of the challenges faced in establishing and maintaining an information security management system on the Brazilian scene
Rodrigo Valle Fazenda (Unisinos), Leonardo Lemes Fagundes (Unisinos)
- 315 An Investigation About the Absence of Validation on Security Quantification Methods
Rodrigo Sanches Miani (UFU), Bruno Bogaz Zarpelão (UEL), Leonardo de Souza Mendes (UNICAMP)

MT9 - Software Requirements, Architecture and Design for Information Systems **323**

- 323 A Creative Process to Elicit Contexts for Context Sensitive Systems
Carlos Batista (UFPE), Carla Silva (UFPE)
- 331 A Risk Management Approach for Software Product Line Engineering
Luanna Lopes Lobato (UFG), Thiago Jabur Bittar (UFG)
- 339 Management of Authorization Rules Using Conceptual Model
Ricardo Diniz Sul (UNIRIO), Bruna Christina P. Brandão (UNIRIO), Leonardo Guerreiro Azevedo (UNIRIO / IBM Research), Fernanda Baião (UNIRIO), Claudia Cappelli (UNIRIO)
- 347 Co-Occurrence of Design Patterns and Bad Smells in Software Systems: An Exploratory Study
Bruno Cardoso (UFMG), Eduardo Figueiredo (UFMG)

MT10 - Transparency **355**

- 355 A Customized Checklist for Analyzing Software Transparency in Websites
Fabio Bittencourt Forte (UFSC), Patrícia Vilain (UFSC), Fabíola Ferreira de Macedo (ALESC)
- 363 Transparent Sharing Architecture of Content Between Mobile Devices in Opportunistic Networks
Charles Tim Batista Garrocho (UFOP), Maurício José da Silva (UFOP), Ricardo Augusto Rabelo Oliveira (UFOP)
- 371 Adoption of open source software: A study on the information technology sector in Minas Gerais
Luciana Guimarães Carvalho (UFMG), Orlando Abreu Gomes (FUMEC), Fernando Silva Parreiras (FUMEC)

MT11 - Applied Technology **379**

- 379 Technological Management of Small Crops Through Mobile Apps and Precision Agriculture
Sergio Manuel Serra da Cruz (UFRRJ), Diogo Nunes (UFRRJ), Carlos Werly (UFRRJ), Pedro Vieira Cruz (UFRRJ), Ana Cláudia de Macedo Vieira (UFRJ), Marden Manuel Marques (CEASA-RJ)
- 387 An Unsupervised Method for Ontology Population from Textual Sources on the Web
Fábio Lima (UFBA), Hilário Oliveira (UFPE), Laís Salvador (UFBA)
- 395 Using Gamification to Improve Treatment Adherence
Matheus Batista Nascimento (UFG), Eduardo Simões de Albuquerque (UFG)

Special Tracks **399**

Experience Reports in Industry and Case Studies (RESI) **399**

- 399 Decision Making in Public Administration Supported by Knowledge Discovery: A Case Study in Project Management
Caue R. do Prado (USP), Sarajane M. Peres (USP), Marcelo Fantinato (USP)
- 407 An analysis Instruments for Assessing Adherence to level 2 TMMi in Small Enterprises
Daniella de Oliveira Costa (UFG), Plínio de Sá Leitão Júnior (UFG), Fábio Nogueira de Lucena (UFG)

- 415 Verbal Anchor Usage on Fuzzy Systems to Help Creating Initial Strategy Maps: Case Study
Marcelo Ladeira (UnB), Fernando de Albuquerque Linhares (UnB)
- 423 Maturity Level Assessment in Software Testing in Small and Medium-Sized Enterprises of the State of Goiás
Adailton F. Araújo (UFG), Cássio L. Rodrigues (UFG), Auri M. R. Vincenzi (UFG), Celso G. Camilo (UFG)
- 431 Evaluation of the Use of Creation Platform of Electronic Medical Records SANA Platform for Basic Health Experts
Taís Bedendo do Valle (UFSJ), Dárlinton Barbosa Feres Carvalho (UFSJ)
- 439 A Case Study on the Usage of the Value Blueprint for Ecosystem Design
Luciana A. Almeida (UFPA), Cleidson R. B. de Souza (ITV, UFPA), Adailton M. Lima (UFPA), Rodrigo Reis (UFPA)
- 447 Restrictions on use of the Factory Floor Information in Maintenance Management
R. J. Kurscheidt Netto (PUC-PR), E. A. P. Santos (PUC-PR), E. de F. R. Loures (PUC-PR)
- 455 Proposal Sentiment Analysis Use in Developing a new Sustainability Metrics
Nataly Bruna Aires (FAMEC), Lucas Vieira Lopes (FAMEC), Eduardo F. Machado (FAMEC), Pietra Sales (FAMEC), Marcos van Vessen Jr (FAMEC), Cassiana F. da Silva (FAMEC)
- 459 Experimentation at Industrial Setting to Improve the Effectiveness of the ETL Procedures Implementation in a Business Intelligence Environment
Juli Kelle Góis Costa (UFS), Igor Peterson Oliveira Santos (UFS), André Vinícius R. P. Nascimento (UFS), Methanias Colaço Júnior (UFS)
- 467 A Systematic Study on Approaches to deal with the Systems Evolution and Customization
Fernanda A. Passos (UFS), Kleber T. O. Santos (UFS), Raphael F. S. Barreto (UFS), Galileu S. de Jesus (UFS), Glayderson J. Nunes (UFS), Lidiany C. Santos (UFS)
- 475 Restructuring Information Technology Area: An Experience Report in the Public Service
Fernando Szimanski (MCTI), Anivaldo S. Vale (MCTI), George H. Kuroki Jr. (MCTI), Célia G. Ralha (UnB)

IS Applications and Tools (ASI)

483

- 483 Minimization of Instructions to Access Memory by Color Flipping in the Interference Graph
Felipe L. Silva (UEL), Marcelo F. Luna (UEL), Wesley Attrot (UEL)
- 487 Open Source Tools Applied to Text Data Recovery in Big Data Environments
Brunno F. M. Attorre (Mackenzie), Leandro A. Silva (Mackenzie)
- 493 An Ontological Approach for Information Management of Public Transport Networks
Ernesto Fonseca Veiga (UFG), Renato de Freitas Bulcão-Neto (UFG)
- 501 M-Health Solution on Pre-diagnosis of Larynx
Estefânia Mayumi Fuzyi (UEL), Matheus Camilo da Silva (UEL), Sylvio Barbon Jr. (UEL)
- 509 Fotosenti: An App to Support on Psychological Treatments
Dárlinton Barbosa Feres Carvalho (UFSJ), Samuel Moreira Abreu Araújo (UFSJ)
- 517 Adding Structured Information to the ANVISA's "Bulário Eletrônico"
João Vitor F. da Silva (UTFPR), Carlos N. Silla Jr. (UTFPR), André Y. Kashiwabara (UTFPR)
- 525 An Approach to Manage Evaluations Using an Interactive Board Game: The Ludo Educational Atlantis
Thiago Jabur Bittar (UFG), Luanna Lopes Lobato (UFG), Leandro Agostini do Amaral (UFG), Elson Longo (UFG)
- 533 Chronos Ações: Tool to Support Decision Making for Investor of the Stock Exchange
Tibério César Souza do Nascimento (UFRN), Leonardo Cunha de Miranda (UFRN)
- 541 RateMyDay: Mobile Tool to Follow the Daily Activities
João Wagner Chaves Costa (UFRN), Leonardo Cunha de Miranda (UFRN)
- 549 A Verbal Anchor based Fuzzy System to help business managers build Balanced Scorecards Strategy Maps
Marcelo Ladeira (UnB), Fernando de Albuquerque Linhares (UnB)
- 555 Knowing the Brazilian Information Systems Community: A Comparative Study with Different Database Approaches
Natan de S. Rodrigues (UnB), Célia G. Ralha (UnB)
- 563 TESE: An Information System for Management of Experimental Software Engineering Projects
Isaque Elcio de Souza (UFG), Paulo Henrique Lima Oliveira (UFG), Esdras Lins Bispo Junior (UFG), Ana Carolina Gondim Inocêncio (UFG), Paulo Afonso Parreira Júnior (UFG)

Information Systems Education (ESI)

571

- 571 Systematic Review on Computer Literacy
Marco Alberto Wang (USP), Edmir Parada Vasques Prado (USP)
- 579 Teaching Information Systems for Computer Graduation Courses: a report experience using a practical ICT approach
Edison Ishikawa (UnB), Célia Ghedini Ralha (UnB)
- 587 An Experiment with the ERP Systems usage to assist the Knowledge Acquisition of Business Processes in Bachelors on Business Administration Course
Márcio Pires de Oliveira (USP), Ildeberto Aparecido Rodello (USP)
- 595 ReqStore: A Software Project Requirements Repository for Supporting PBL Methodology
Paulo I. A. Pordeus (UFC), Carla I. M. Bezerra (UFC), Emanuel F. Coutinho (UFC)
- 603 Qualification and Acknowledgement of Information Systems Professionals
Eliane Cristina de Freitas Rocha (UFMG)
- 611 Gamified Courses in Information Systems Program
Karen da Silva Figueiredo (UFMT)

Government Information Systems (SIG)

615

- 615 The Propensity Of Users To Technologies Adoption: A Study With Users and Not Users of the “Nota Legal” Program In The Federal District
Josivania Silva Farias (UnB), Paula Valente Lins (UnB), Pedro H. M. Albuquerque (UnB)
- 623 Study About User’s Engagement of a Governmental Social Media
Camila Mariane Costa Silva (USP), Edmir Parada Vasques Prado (USP)
- 631 ArgumentBind - A Model for Implementing Argument Web Integrated Applications with Open and Linked Data
Roberto Niche (Unisinos), Sandro J. Rigo (Unisinos)
- 639 SIGS-S Mobile Saúde: A Mobile Application to Support the Collection of Health Data
Eduardo Fernandes (UFMS), Marcelo Turine (UFMS), Maria Istela Cagnin (UFMS)
- 647 Inspection of Governmental Sites Interaction: Comparison Between Methods
Vinícius de Figueiredo Marques (UFMT), Breno Augusto Ferreira (UFMT), João Paulo Pinho (UFMT), Rondinely Oliveira (UFMT), Cristiano Maciel (UFMT)
- 655 Digital Ecosystems to Support the Open and Collaborative Government Systems
Andréa Magalhães Magdaleno (UNIRIO), Renata Mendes de Araujo (UNIRIO)
- 659 Integrated System of Municipal Costs: a Tool to Support Decision Making for Public Manager
Anderson P. Avila-Santos (UEL), Daniel S. Kaster (UEL), Evandro Baccarin (UEL), Letícia F. Negreiros (UEL), Saulo F. Amâncio Vieira (UEL)
- 667 Comparing Text Mining Algorithms for Predicting Irregularities in Public Accounts
Breno Santana Santos (UFS), Methanias Colaço Júnior (UFS), Bruno Cruz da Paixão (UEL), Rafael M. Santos (UEL), André Vinícius R. P. Nascimento (UEL), Hallan Cosmo dos Santos (TCE-SE), Wallace H. L. Filho (TCE-SE), Arquimedes S. L. de Medeiros (TCE-SE)
- 675 Discussing a Recommendation System to Support Participation in the Democratic Citizen Community
Max Martins (AL-MT), Jivago Medeiros (UFMT), Cristiano Maciel (UFMT)
- 683 Statistical Audit via Gaussian Mixture Models in Business Intelligence Systems
Bruno H. A. Pilon (UnB), João Paulo C. L. da Costa (UnB), Juan J. Murillo-Fuentes (University of Sevilla), Rafael T. de Sousa Júnior (UnB)
- 691 Accessibility in Electronic Government: A Study on the Implementation of Web Standards in Sites gov.br
Alberto Dumont Alves Oliveira (USP), Marcelo Medeiros Eler (USP)
- 699 SmartCluster: Using Public Data to Group Smart Cities by Domains
Ricardo Alexandre Afonso (UFAL), Clóvis Holanda do Nascimento (UFPE), Vinicius Cardoso Garcia (UFPE), Alexandre Álvaro (UFSCAR)
- 703 Inferring Companies Similarities from Brazilian Government Expenditure Data
Marcelo Pita (SERPRO), Gustavo da Gama Torres (SERPRO)

Business Process Management (BPM)

709

- 709 Analysis of Profile and Role from Business Analysts in National Context
Viviane Cortese Zadra (Unisinos), Josiane Brietzke Porto (Unisinos)
- 717 BPM Maturity Models: A Qualitative Study under the Perspective of Specialists
Simone Nunes da Nóbrega (UFPB), Joyce Aline Oliveira (UFPE), Rodrigo Rebouças (UFPB)

- 725 Major Issues in Business Process Management: Key Concerns presented in Academy from a Brazilian
Valdemar T. F. Confort (UNIRIO), Flávia Maria Santoro (UNIRIO)
- 733 How Organizational Culture influences BPM Evolution
Iveruska Jatobá (UFPE), Carina Alves (UFPE)
- 741 Towards Aspects Identification in Business Process Through Process Mining
Bruna Christina P. Brandão (UNIRIO), Flávia Maria Santoro (UNIRIO), Leonardo Guerreiro Azevedo (IBM Research)
- 749 A Method to Identify Variability in Knowledge Intensive Processes
Anderson S. do Nascimento (UNIRIO), Flávia Maria Santoro (UNIRIO), José Ricardo Cereja (UNIRIO)

Index of Authors

757

Sistema para roteamento de veículos capacitados aplicando *Métodos de Monte Carlo*

Alternative Title: Capacitated vehicle routing system applying Monte Carlo methods

Rômulo Augusto de Carvalho Oliveira
Universidade de São Paulo
romulo.augusto.oliveira@usp.br

Karina Valdivia Delgado
Universidade de São Paulo
kvd@usp.br

RESUMO

O Problema de Roteamento de Veículos (*Vehicle Routing Problem*, VRP) é um dos problemas de Otimização Combinatória mais estudados dentro da Computação e de grande relevância para as áreas de logística e transporte. Este trabalho apresenta um novo algoritmo para resolução do Problema de Roteamento de Veículos Capacitados (*Capacitated Vehicle Routing Problem*, CVRP), utilizando Métodos de Monte Carlo. Métodos de Monte Carlo são métodos estatísticos que utilizam amostragem aleatória para resolver problemas probabilísticos e determinísticos. O algoritmo proposto foi desenvolvido baseado em Simulações de Monte Carlo e na heurística de *Clarke & Wright Savings* e demonstrou resultados comparáveis aos melhores algoritmos existentes na literatura, superando trabalhos anteriores com Métodos de Monte Carlo. A comparação, análise e avaliação do algoritmo foram feitas com base em *benchmarks* de problemas existentes na literatura.

Palavras-Chave

Problema de Roteamento de Veículos Capacitados, Métodos de Monte Carlo, Otimização Combinatória, Algoritmos Heurísticos, Algoritmos Meta-heurísticos

ABSTRACT

The Vehicle Routing Problem (VRP) is one of the combinatorial optimization problems most studied in Computer Science and of great relevance to the areas of logistics and transport. This paper presents a new algorithm for solving the Capacitated Vehicle Routing Problem (CVRP) using Monte Carlo methods. Monte Carlo methods are statistical methods that use random sampling to solve probabilistic and deterministic problems. The proposed algorithm was developed based on Monte Carlo simulations and Clarke and Wright Savings heuristic and demonstrated results comparable to the best existing algorithms in the literature, it

overcomes previous work with Monte Carlo methods. The comparison, analysis and evaluation of the algorithm were based on existing benchmarks in the literature.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous;
H.4.2 [Types of Systems]: [Logistics]

General Terms

Theory, Algorithms

Keywords

Capacitated Vehicle Routing Problem, Monte Carlo Methods, Combinatorial Optimization, Heuristic Algorithms, Meta-heuristic Algorithms

1. INTRODUÇÃO

O Problema de Roteamento de Veículos (*Vehicle Routing Problem*, VRP) é um problema de otimização combinatória que vem sendo estudado na área de Computação há mais de cinquenta anos [12]. É uma variação do Problema do Caixeiro Viajante [17, 18], portanto da classe NP-difícil [21], e resume-se em atender um número determinado de clientes através de uma frota de veículos, utilizando a rota que ofereça o menor custo possível.

Além de ser um problema importante de otimização combinatória muito estudado em Computação [32], o roteamento de veículos é parte crítica e fundamental para execução bem-sucedida do processo de logística e transporte [16]. Em uma sociedade onde a eficiência energética e a redução de poluentes são preocupações globais, o roteamento de veículos feito de forma eficiente pode reduzir custos, economizar energia e reduzir a emissão de gases poluentes [9].

Neste trabalho, será abordada a variação mais conhecida [29] do VRP: O Problema de Roteamento de Veículos Capacitados (*Capacitated Vehicle Routing Problem*, CVRP). O CVRP consiste no Problema de Roteamento de Veículos com a restrição de que cada veículo possui uma capacidade máxima de carga. Métodos de Monte Carlo serão aplicados para resolver o CVRP. Métodos de Monte Carlo são métodos estatísticos que utilizam amostragem aleatória para resolver problemas probabilísticos e determinísticos [7]. Essas técnicas são aplicadas quando não é viável empregar métodos exatos na resolução de problemas. O algoritmo proposto foi

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

desenvolvido baseado em Simulações de Monte Carlo e na heurística de *Clarke & Wright Savings*.

A Seção 2 apresenta a definição formal de CVRP. A Seção 3 introduz Métodos de Monte Carlo. A Seção 4 discute sobre algoritmos utilizados para resolver CVRP. A aplicação dos Métodos de Monte Carlo em CVRP e o algoritmo proposto neste trabalho são discutidos na Seção 5. Na Seção 6 são apresentadas a metodologia e os experimentos realizados. A Seção 7 discute sobre os resultados obtidos e a Seção 8 conclui este trabalho.

2. DEFINIÇÃO FORMAL DE CVRP

O Problema de Roteamento de Veículos Capacitados pode ser formalizado como um grafo não direcionado $G = (V, E)$, onde $V = \{v_0, v_1, \dots, v_n | n \geq 1\}$ é o conjunto de vértices e $E = \{(v_i, v_j) : v_i, v_j \in V, v_i \neq v_j\}$ é o conjunto de arestas. Definimos m como o número de veículos a serem usados para servir todos os clientes, onde $1 \leq m \leq n$. Os veículos têm a mesma capacidade máxima Q . O vértice v_0 representa o depósito, de onde partem os veículos, e cada um dos demais vértices representa um cliente. Cada vértice $v_i \in V$ tem associada uma demanda q_i , onde $0 < q_i \leq Q$. Cada aresta $(v_i, v_j) \in E$ tem associada um custo não negativo c_{ij} , que representa o custo de se chegar do vértice v_i ao vértice v_j .

A solução consiste em encontrar um conjunto de rotas $R = \{r_1, r_2, \dots, r_m\}$, de forma que todos os clientes sejam visitados exatamente uma vez, sem que as rotas excedam a capacidade máxima Q dos veículos e com o menor custo $C(R)$ possível. Cada rota r_i é servida por exatamente um veículo, começa e termina no depósito, tem tamanho k_i e é definida por $r_i = (v_0^i, v_1^i, \dots, v_{k_i}^i, v_{k_i+1}^i)$, onde $v_0^i = v_{k_i+1}^i = v_0$, $v_j^i \neq v_l^i$ para $0 \leq j < l \leq k_i$. Cada rota r_i tem um custo $C(r_i) = \sum_{j=0}^{k_i} c_{v_j^i, v_{j+1}^i}$, que é o custo de percorrer a rota completamente (partindo do depósito, passando por todos os clientes e retornando ao depósito). O custo da solução $C(R)$ é a soma do custo das rotas, definido como $C(R) = \sum_{i=1}^m C(r_i)$.

O CVRP consiste em um problema de minimização e a melhor solução é aquela que apresenta o menor custo $C(R)$.

3. MÉTODOS DE MONTE CARLO

Métodos de Monte Carlo são uma classe de métodos que utilizam amostragem de números aleatórios para se obter resultados numéricos, isto é, utilizam métodos probabilísticos para resolver problemas probabilísticos e determinísticos [7]. Definido em 1949 por Nicholas Metropolis e Stanislaw Ulam, foi desenvolvido por John Von Neumann e Stanislaw Ulam [23].

Simulação de Monte Carlo (*Monte Carlo Simulation*, MCS) é uma das técnicas usadas em análise probabilística para se entender distribuições e extrair seu comportamento de amostras aleatórias. Essas amostras são geradas a fim de permitir que o comportamento real possa ser simulado, estudado e entendido sem que novas amostras populacionais precisem ser colhidas [24]. O Método de Monte Carlo é muito utilizado, dentre outras coisas, para integrações em muitas dimensões, resolver sistemas lineares densos e simular modelos físicos e matemáticos complexos [15].

Amostragem aleatória, em inglês *random sampling*, é o processo estatístico de seleção de uma amostra populacional. O tipo de amostragem aleatória abordado neste trabalho é

a amostragem aleatória simples (*plain random sampling* ou *simple random sampling*) [8].

Amostragem aleatória simples é um tipo de amostragem probabilística onde os elementos da população são sorteados aleatoriamente. Todos os elementos têm a mesma probabilidade de serem selecionados. O sorteio é feito através de números aleatórios gerados por computadores ou por numeração e sorteio usando tabelas de números aleatórios [8, 11]. Em problemas de busca, *plain random sampling* pode ser usado como um método que começa no nó raiz e continuamente escolhe nós filhos aleatórios até que um nó folha seja alcançado. A cada nível da árvore, um nó é sorteado e escolhido para continuar com a busca. Esse método pode ser aplicado a fim de maximizar a diversidade das soluções encontradas quando o espaço de busca é muito grande para usar a busca em profundidade [30].

Plain random sampling pode ser aplicado em Simulações de Monte Carlo para problemas de busca [30]. Primeiro é selecionado o nó raiz na árvore de busca. Para cada um dos n nós filhos candidatos, são executadas r simulações aleatórias até que um nó folha seja alcançado. A simulação é feita utilizando o método *plain random sampling*, explicado anteriormente. Após esse processo, o melhor nó filho candidato é escolhido entre os nós existentes para fazer parte da solução. O melhor nó pode ser o que obteve melhor solução entre todas as simulações ou o que obteve a melhor média entre as simulações [30]. A partir daí, esse nó filho passa ser o nó atual da árvore de busca e o método continua até que o nó atual seja uma folha. A vantagem da Simulação de Monte Carlo é que a busca é guiada de acordo com o melhor resultado obtido dentre as simulações. Esse processo pode ser repetido várias vezes para se obter um conjunto de soluções ou a melhor solução entre todas as iterações.

4. ALGORITMOS DE CVRP

Há duas abordagens principais usadas para resolver o problema de roteamento de veículos: métodos exatos e métodos heurísticos. Os métodos exatos são soluções que garantem que a solução ótima será encontrada, mas que pode ser custoso computacionalmente devido à complexidade NP-difícil [20]. Entre os algoritmos exatos destacam-se *branch and bound* [19, 14] e *branch and cut* [22].

Métodos heurísticos são métodos que percorrem um caminho reduzido no espaço de busca seguindo uma estratégia guiada e não garantem que a solução ótima será encontrada, mas permitem encontrar uma aproximação com menor custo de tempo e recursos [26]. Os métodos heurísticos são divididos em duas classes principais: métodos heurísticos clássicos e meta-heurísticos [32].

Algoritmos heurísticos clássicos geralmente encontram soluções boas em tempo reduzido e sem explorar profundamente o espaço de busca [32]. Entre os métodos heurísticos clássicos destaca-se o *Clarke & Wright Savings* [10]. Esse método será abordado neste trabalho e usado como algoritmo base para aplicação dos Métodos de Monte Carlo.

Métodos meta-heurísticos geralmente incorporam métodos heurísticos clássicos, mas fazem uma exploração maior de regiões mais promissoras do espaço de busca. Esses métodos geralmente retornam soluções melhores do que os métodos tradicionais, ao custo de tempo de computação [32]. Algoritmos meta-heurísticos utilizam conceitos de diversas áreas para desenvolver heurísticas que possam obter melhores soluções dos problemas [25]. Entre as áreas podem ser

destacadas Inteligência Artificial [4, 6] e Matemática [27].

4.1 Algoritmo Clarke & Wright Savings

O algoritmo *Clarke & Wright Savings* (CWS) é um dos algoritmos heurísticos mais conhecidos e estudados em CVRP [32]. Foi definido em 1964 [10] e é baseado no princípio de economia (*savings*). É um método construtivo, ou seja, a solução vai sendo construída a cada passo, sempre se preocupando com as restrições do problema. O princípio do algoritmo se baseia na ideia de que se há duas rotas diferentes $r = (0, \dots, i, 0)$ e $s = (0, j, \dots, 0)$ que são factíveis de uma fusão (*merge*) em uma nova rota $t = (0, \dots, i, j, \dots, 0)$, a fusão dessas rotas gera uma economia de custo S_{ij} [10], definida por:

$$S_{ij} = C(0, i) + C(0, j) - C(i, j). \quad (1)$$

Duas rotas r e s são factíveis de uma fusão quando a soma das demandas dos clientes das duas rotas não excede a capacidade máxima Q de um veículo:

$$\sum_{i \in r} q_i + \sum_{j \in s} q_j \leq Q. \quad (2)$$

Existem duas versões do algoritmo *Clarke & Wright Savings*: a versão paralela e a versão sequencial. A diferença entre as versões está na estratégia de seleção e no número de rotas construídas a cada passo do algoritmo. Enquanto a versão sequencial apenas trata de uma rota por vez, a versão paralela permite que mais de uma rota seja construída. A versão paralela do algoritmo supera os resultados da versão sequencial [32] e será abordada neste trabalho. O Algoritmo 1 contém o pseudocódigo do algoritmo paralelo.

Algorithm 1: Clarke & Wright Savings

Entrada: Conjunto de vértices V e conjunto de arestas E , capacidade máxima Q dos veículos

Saída: Conjunto de rotas

```

1 routes ← create_initial_routes(V);
2 savings_list = compute_savings_list(V);
3 foreach  $i, j \in \text{savings\_list}$  do
4   if feasible_merge( $i, j, Q$ ) then
5     routes ← merge_routes( $i, j$ );
6 return routes
```

O Algoritmo 1 começa colocando cada cliente i em uma nova rota $r = (0, i, 0)$ que vai até o cliente e retorna para o depósito (Linha 1 do Algoritmo 1). Depois, é criada a lista de *savings* (*savings list*) com a economia obtida entre cada nó $i, j \in V - \{0\}$, ordenada de forma decrescente de economia. O Algoritmo 2 contém o pseudocódigo do algoritmo de criação da lista de savings.

Nos passos das Linhas 3-5 do Algoritmo 1, a lista de *savings* é iterada e para cada par (i, j) o Algoritmo CWS tenta fundir as rotas as quais i e j pertencem (Linha 4). A Linha 4 verifica as condições para que duas rotas sejam consideradas factíveis de fusão, já citadas anteriormente:

1. i deve ser o primeiro (ou último) cliente a ser visitado na sua rota r ,
2. j deve ser o último (ou o primeiro) cliente a ser visitado na sua rota s ,

Algorithm 2: compute_savings_list

Entrada: Conjunto de vértices V

Saída: Lista de *savings*, ordenada por ordem decrescente de economia

```

1 savings_list = {};
2 foreach  $i, j \in V - \{0\}$  do
3    $S_{ij} = C(0, i) + C(0, j) - C(i, j)$ ;
4   savings_list ←  $S_{i,j}$ ;
5 sort_decreasing(savings_list);
6 return savings_list
```

3. r deve ser diferente de s e

4. a capacidade das rotas juntas não deve exceder Q (ver Equação 2).

Quando duas rotas r e s podem ser fundidas, uma nova rota t ligando as rotas r e s é criada. As rotas r e s são eliminadas e a rota t é adicionada na lista de rotas (Linha 5). No final, a lista de rotas criada (*routes*) é retornada pelo algoritmo.

4.2 Algoritmos utilizando Métodos de Monte Carlo

Comparado com a literatura existente para o Problema de Roteamento de Veículos, a aplicação dos Métodos de Monte Carlo em VRP e CVRP é relativamente pequena [30]. O algoritmo *ALGACEA-1* foi proposto em [13] utilizando Técnicas de Monte Carlo sobre o algoritmo de Clarke & Wright Savings. Logo depois, os autores propuseram o *ALGACEA-2*, uma versão melhorada do *ALGACEA-1* que introduziu no algoritmo o conceito de Entropia [13, 31]. Em [29], foram propostas diversas formas de aplicar MCS em problemas de roteamento de veículos, entre elas um algoritmo baseado em *Nearest Neighbor Insertion* e dois algoritmos baseados em CWS, o *BestX-CWS* e o *BinaryMCS-CWS*. O algoritmo *BinaryMCS-CWS* supera os resultados de todos os trabalhos anteriores com MCS com resultados comparáveis aos melhores algoritmos de CVRP existentes na literatura [29].

O algoritmo *BinaryMCS-CWS* trabalha percorrendo a lista de *savings* do CWS e então efetuando um número determinado de simulações. A estratégia usada no algoritmo é explorar a árvore de busca de forma binária, isto é, para cada par de nós (i, j) são feitas r simulações com o par *pertencendo* à solução e r simulações *sem o par pertencer* à solução. O caminho que obtiver o melhor custo médio determina se o par deve ou não ser incluído na solução que está sendo construída [29].

Em cada simulação, um par de nós (i, j) escolhido da lista de *savings* só é processado com uma probabilidade p , com $0 \leq p \leq 1$. O algoritmo pula alguns pares de nós durante as simulações e gera soluções diversificadas. Valores altos de p fazem com que muitos pares sejam pulados, gerando certo “caos”, enquanto valores menores não permitem tanta exploração do espaço de busca [29]. O autor demonstra $0.05 \leq p \leq 0.4$ como o intervalo de p que gera soluções melhores para o algoritmo.

O algoritmo *BinaryMCS-CWS*, portanto, explora a árvore de busca e constrói uma solução de acordo com os resultados obtidos dentro das simulações. Isso guia a estratégia para o ramo de melhores resultados. Entretanto, a melhor

solução pode ser encontrada em um ramo da árvore de busca percorrido durante uma simulação e não necessariamente ao término da iteração da lista de *savings* e da solução construída.

5. MONTE CARLO SAVINGS

Nesta seção, um novo algoritmo baseado no algoritmo *Clarke & Wright Savings* é proposto usando técnicas de Simulação de Monte Carlo. O algoritmo foi inspirado por aplicações anteriores dos Métodos de Monte Carlos e melhorias propostas no algoritmo de *CWS*.

Em todos os algoritmos que utilizam o algoritmo de *Clarke & Wright Savings*, as técnicas de Monte Carlo são aplicadas para aumentar a variabilidade das soluções a fim de superar os limites da heurística utilizada. Todos os algoritmos utilizam *MCS* durante a etapa de seleção de nós/arestas e fusão de rotas do algoritmo *CWS*.

Neste trabalho, uma abordagem diferente será feita utilizando a heurística de *Clarke & Wright Savings*. Simulações de Monte Carlo serão usadas para aumentar a variabilidade da lista de *savings*. A ideia é manter o algoritmo original de construção de soluções, mas utilizando as simulações para variar a ordem em que as arestas são percorridas, mudando as soluções finais e permitindo que diferentes rotas sejam geradas em cada simulação.

5.1 Algoritmo Monte Carlo Savings

O algoritmo, batizado de *Monte Carlo Savings*, é um algoritmo heurístico novo baseado no *Clarke & Wright Savings* [10], que utiliza Simulações de Monte Carlo para gerar lista de *savings* com certo grau de variabilidade e permitir que soluções diversificadas sejam encontradas. O algoritmo *Monte Carlo Savings* (Algoritmo 3) executa r simulações, onde em cada uma delas é gerada uma lista de *savings* utilizando *plain random sampling*. Para cada lista gerada, o algoritmo procede ao método padrão de *Clarke & Wright Savings*, percorrendo a lista de forma decrescente de economia e fundindo as rotas factíveis. Para permitir a variabilidade nas listas de *savings*, a fórmula de cálculo de economia recebe um componente aleatório p , definido em um intervalo $[-\lambda, +\lambda]$, usado para penalizar ou bonificar a economia gerada entre dois nós i e j . Uma economia é penalizada quando $-\lambda \leq p < 0$ e bonificada quando $0 < p \leq \lambda$. Quando $p = 0$, a economia não tem alteração. Seja $s = C(0, i) + C(0, j) - C(i, j)$, a Equação 3 mostra o cálculo da nova economia entre dois clientes i e j .

$$S_{ij} = s + s \cdot p, \quad (3)$$

em que $-\lambda \leq p \leq \lambda$. Essas alterações na economia entre dois clientes permitem que a lista de *savings* tenha as arestas ordenadas de forma diferente do método padrão, fazendo com que a árvore de busca seja mais explorada e soluções diferentes sejam encontradas.

Na Linha 1 do Algoritmo 3, é criada uma solução base com cada cliente sendo servido por seu próprio veículo. A função *create_initial_routes* é a mesma usada pelo *CWS* no Algoritmo 1. Nas Linhas 2-9 do Algoritmo 3, são efetuadas r simulações. Em cada simulação, é criada uma solução base (Linha 3), computada a lista de *savings* (Linha 4) usando o algoritmo *compute_mcs_savings* (Algoritmo 4), e calculada as rotas através do método clássico de *Clarke & Wright Savings* (Linhas 5-7). A função *merge_routes* é a mesma

Algorithm 3: Monte Carlo Savings

Entrada: Conjunto de vértices V e conjunto de arestas E , capacidade máxima Q dos veículos, número de simulações r e limite do intervalo aleatório λ

Saída: Conjunto de rotas

```

1 best = create_initial_routes(V);
2 for  $k \leftarrow 0$  to  $r$  do
3   solution = create_initial_routes(V);
4   savings_list = compute_mcs_savings(V,  $\lambda$ );
5   foreach  $i, j \in \text{savings\_list}$  do
6     if feasible_merge( $i, j, Q$ ) then
7       solution  $\leftarrow$  merge_routes( $i, j$ ); // CWS
8   if  $C(\text{solution}) < C(\text{best})$  then
9     best = solution;
10 return best
```

usada pelo *CWS* definida no Algoritmo 1. Finalmente, nas Linhas 8-9 do Algoritmo 3, compara-se o custo da solução (definido na seção 2) da simulação atual com o custo da melhor solução encontrada até agora e a solução com menor custo é escolhida como a melhor (*best*). O algoritmo termina quando r simulações são executadas e a melhor solução encontrada dentre as simulações é retornada. O Algoritmo 4 mostra a criação da lista de *savings*, utilizando o parâmetro λ para definir o intervalo em que o valor da componente aleatória p pode estar.

Algorithm 4: compute_mcs_savings

Entrada: Conjunto de vértices V e limite do intervalo aleatório λ

Saída: Lista de *savings*, ordenada por ordem decrescente de economia

```

1 savings_list = {};
2 foreach  $i, j \in V - \{0\}$  do
3    $s = C(0, i) + C(0, j) - C(i, j)$ ;
4    $p = \text{random}(-\lambda, +\lambda)$ ; // random  $p \in [-\lambda, +\lambda]$ 
5    $S_{i, j} = s + (s \cdot p)$ ;
6   savings_list  $\leftarrow S_{i, j}$ ;
7 sort_decreasing(savings_list);
8 return savings_list
```

Note que para $r = 1$ e $\lambda = 0$, o algoritmo *Monte Carlo Savings* atua como o algoritmo padrão de *Clarke & Wright Savings*.

5.2 Análise de complexidade assintótica

Para analisarmos a complexidade assintótica do algoritmo *Monte Carlo Savings* é preciso primeiro analisar a complexidade do *CWS*, usado como base pelo nosso algoritmo. O consumo de tempo do algoritmo *Clarke & Wright Savings*, pode ser representada pela função $f(n, m)$, onde n é o número de clientes do problema e m o número de arestas. Essa função é obtida através da análise de complexidade dos 4 passos principais do algoritmo: inicialização das rotas, criação da lista de *savings*, ordenação da lista de *savings* e fusão das rotas. O primeiro passo, inicialização das rotas (*create_initial_routes* do Algoritmo 1) tem complexidade linear n . Os passos de

criação e ordenação da lista de *savings* (*compute_savings_list* do Algoritmo 1) têm complexidade $m + m \cdot \log_2(m)$. Os passos de fusão das rotas, representado nas Linhas 3-5 do Algoritmo 1 também têm complexidade linear m . Assim, podemos concluir que o consumo de tempo do algoritmo *Clarke & Wright Savings* é $f(m, n) = n + m + m \cdot \log_2(m) + m$.

Para facilitar, podemos tomar $m = n^2$, embora o número máximo de arestas em um problema de CVRP seja de $n \cdot (n-1)/2$. Portanto, o consumo de tempo do algoritmo CWS é $O(n^2 \cdot \log_2(n))$.

O algoritmo *Monte Carlo Savings*, de forma simplificada, executa r vezes o algoritmo de *CWS* usando números aleatórios para variar na criação da lista de *savings*. Note que a função *compute_mcs_savings* também tem a mesma complexidade da função *compute_savings_list* do Algoritmo 1. Portanto, o consumo de tempo do algoritmo de *Monte Carlo Savings* é $O(r \cdot n^2 \cdot \log_2(n))$.

É importante ressaltar que o número de simulações r pode ser muito maior do que o número de clientes n .

6. EXPERIMENTOS

O algoritmo proposto é comparado com algoritmos de CVRP heurísticos, meta-heurísticos e que usam dos Métodos de Monte Carlo, entre eles os algoritmos heurísticos *Clarke & Wright Savings* [10] e baseado em *Centroide* [28, 5]; o algoritmo meta-heurístico *Busca Tabu Granular* [33, 5]; e o algoritmo que aplica Métodos de Monte Carlo *BinaryMCS-CWS* [29].

Para avaliar e comparar o desempenho dos algoritmos, foram usados um subconjunto de 15 instâncias de problemas definidos por Augerat [3] e um subconjunto de 13 instâncias de problemas utilizados por Takes [30] para avaliar o desempenho do seu algoritmo *BinaryMCS-CWS* frente ao *ALGACEA-2*. O conjunto de problemas e soluções dos *benchmarks* utilizados está no formato TSPLIB [1].

O *benchmark* de Augerat foi usado para comparar os algoritmos de *Centroide*, *Busca Tabu Granular*, *Clarke & Wright Savings*, *BinaryMCS-CWS* e o *Monte Carlo Savings*. O *benchmark* de Takes foi usado para comparar os algoritmos *BinaryMCS-CWS* e o algoritmo proposto *Monte Carlo Savings*.

Foram avaliados os resultados baseados na melhor resposta obtida pelos algoritmos para aquela determinada instância de problema. As comparações também levam em conta a solução ótima disponível na literatura. Para este trabalho, o algoritmo proposto *Monte Carlo Savings* foi implementado usando a linguagem de programação *Python*. Também foi implementada a versão paralela do algoritmo de *Clarke & Wright Savings* e o algoritmo *BinaryMCS-CWS*. O código fonte dessas implementações está disponível em <https://github.com/RomuloOliveira/monte-carlo-cvrp/>.

Foram feitas 2000 simulações em cada execução do algoritmo *Monte Carlo Savings*. O algoritmo foi executado cinco vezes para cada instância e o melhor resultado obtido foi selecionado. O tempo limite máximo estipulado por execução de instância foi de 5 minutos, o mesmo tempo usado por [30] e [13]. Os resultados da *Busca Tabu Granular* e do algoritmo de *Centroide* são os mesmos mostrados em [5]. Os resultados do *BinaryMCS-CWS* e do *CWS* foram obtidos através de implementação próprias. Foram feitas 50 simulações por cada aresta visitada no algoritmo *BinaryMCS-CWS*, utilizando os mesmos parâmetros de [29]. O algoritmo também foi executado cinco vezes para cada instância

e o seu melhor resultado selecionado.

Os testes foram executados em um computador com processador *AMD Phenom II X4 840 @ 3,2 GHz* e 4GB de memória RAM.

6.1 Benchmark Augerat

Foram utilizadas 15 instâncias do *benchmark* Augerat [3] para comparação dos resultados. As instâncias escolhidas são as mesmas utilizadas por [5]. O número de clientes por instância varia de 16 a 66 clientes.

A Tabela 1 mostra a comparação de resultados entre os algoritmos e a solução ótima existente na literatura para essas instâncias.

Nota-se que o algoritmo *Monte Carlo Savings* obtém os melhores resultados entre os algoritmos testados em 9 das 15 instâncias, incluindo 2 soluções ótimas e também o melhor resultado geral, com apenas 1,39% de variação da solução ótima. Nas instâncias *P-n16-k8* e *P-n23-k8* o algoritmo de *Clarke & Wright Savings* não obteve uma solução factível com o número mínimo desejado de veículos. Foi usado $\lambda = 0,05$ para 13 instâncias, mostrando ser esse um bom valor para o parâmetro. Na instância *B-n31-k5* foi usado $\lambda = 0,0025$ e na instância *P-n19-k2* foi usado $\lambda = 0,1$.

Os resultados de ambos os algoritmos que utilizam Simulações de Monte Carlo mostram-se melhores que as outras abordagens, inclusive a *Busca Tabu Granular*.

A Tabela 2 mostra o resumo do desempenho dos algoritmos. O *Monte Carlo Savings* obteve melhor desempenho considerando o melhor resultado e a média de resultados.

Tabela 2: Resumo do desempenho dos algoritmos para as instâncias do Augerat

Algoritmos	Melhor resultado (%)	Média dos resultados (%)	Pior resultado (%)
Monte Carlo Savings	0,00%	1,19%	3,95%
BinaryMCS-CWS	0,00%	1,32%	3,19%
Busca Tabu	0,00%	3,54%	8,32%
CWS	0,74%	6,36%	18,60%
Centroide	3,64%	11,19%	29,91%

6.2 Benchmark Takes

Foram utilizadas 13 instâncias, sendo duas de [3] e 11 de [2]. Como mencionado anteriormente, as instâncias são as mesmas utilizadas por [30] em sua comparação com o algoritmo *ALGACEA-2*. O número de clientes por instância varia de 51 a 200 clientes para esse subconjunto.

A Tabela 3 mostra a comparação de resultados entre os algoritmos e a solução ótima existente na literatura para essas instâncias.

Novamente, o algoritmo *Monte Carlo Savings* obteve os melhores resultados, agora em 9 das 13 instâncias, com variação de 3,13% da solução ótima. Foi usado $\lambda = 0,0025$ para 10 instâncias, mostrando-se um bom valor para o parâmetro para instâncias maiores, em contrapartida de $\lambda = 0,05$ como melhor valor para o parâmetro do *benchmark* anterior. Nas instâncias *E101-14U* e *E200-17B* foi usado $\lambda = 0,05$ quanto na instância *E051-05E* foi usado $\lambda = 0,8$.

A Tabela 4 mostra o resumo do desempenho dos dois algoritmos.

TabelaTable 1: Comparação de resultados para as instâncias do Augerat

Instância	Solução ótima	Busca Tabu		Centroide		CWS		BinaryMCS-CWS		Monte Carlo Savings	
		Resultado	Diferença	Resultado	Diferença	Resultado	Diferença	Resultado	Diferença	Resultado	Diferença
A-n32-k5	784	784	0,00%	874	11,48%	834	6,38%	796	1,53%	787	0,38%
A-n33-k6	742	760	2,43%	769	3,64%	880	18,60%	744	0,27%	742	0,00%
A-n36-k5	799	829	3,75%	899	12,52%	806	0,88%	805	0,75%	805	0,75%
A-n45-k7	1146	1197	4,45%	1257	9,69%	1203	4,97%	1178	2,79%	1155	0,79%
A-n63-k10	1314	1355	3,12%	1476	12,33%	1382	5,18%	1342	2,13%	1341	2,05%
B-n31-k5	672	675	0,45%	700	4,17%	677	0,74%	674	0,30%	674	0,30%
B-n34-k5	788	790	0,25%	976	23,86%	806	2,28%	793	0,63%	795	0,89%
B-n38-k6	805	821	1,99%	835	3,73%	834	3,60%	809	0,50%	821	1,99%
B-n44-k7	909	946	4,07%	963	5,94%	939	3,30%	926	1,87%	926	1,87%
B-n66-k9	1316	1423	8,13%	1420	7,90%	1424	8,21%	1358	3,19%	1368	3,95%
P-n16-k8	450	452	0,44%	478	6,22%	(478)	6,22%	450	0,00%	450	0,00%
P-n19-k2	212	223	5,19%	242	14,15%	240	13,21%	212	0,00%	215	1,42%
P-n23-k8	529	573	8,32%	595	12,48%	(537)	1,51%	529	0,00%	533	0,76%
P-n40-k5	458	490	6,99%	595	29,91%	522	13,97%	470	2,62%	464	1,31%
P-n50-k10	696	720	3,45%	765	9,91%	740	6,32%	718	3,16%	706	1,44%
Total	11620	12038	3,6%	12844	10,53%	12302	5,87%	11804	1,58%	11782	1,39%

TabelaTable 3: Comparação de resultados para as instâncias usadas por Takes

Instância	Solução ótima	BinaryMCS-CWS		Monte Carlo Savings	
		Resultado	Diferença	Resultado	Diferença
A-n65-k9	1174	1224	4.26%	1197	1,96%
A-n80-k10	1764	1805	2.32%	1812	2,72%
E051-05E	525	536	2.10%	535	1,90%
E072-04F	242	265	9.50%	254	4,96%
E076-07S	691	703	1.74%	701	1,45%
E076-10E	837	860	2.75%	859	2,63%
E076-14U	1029	1057	2.72%	1068	3,79%
E101-08E	826	861	4.24%	859	4,00%
E101-10C	820	844	2.93%	832	1,46%
E101-14U	1091	1101	0.92%	1105	1,28%
E151-12C	1031	1084	5.14%	1084	5,14%
E200-17B	1291	1346	4.26%	1364	5,65%
E200-17C	1311	1360	3.74%	1358	3,59%
Total	12632	13046	3.28%	13028	3.13%

Tabela 4: Resumo do desempenho dos algoritmos para as instâncias usadas por Takes

Algoritmos	Melhor resultado (%)	Média dos resultados (%)	Pior resultado (%)
Monte Carlo Savings	1.28%	3.12%	5.65%
BinaryMCS-CWS	0.92%	3.59%	9.50%

O *Monte Carlo Savings* mostrou bom desempenho também em instâncias com grande número de clientes. Porém, nesse segundo *benchmark*, a diferença entre os algoritmos é menor. O resultado mais próximo da solução ótima foi obtido pelo *BinaryMCS-CWS*, enquanto o *Monte Carlo Savings* foi melhor nos outros dois quesitos (média dos resultados e pior resultado), além do resultado geral.

7. DISCUSSÃO

A aplicação das técnicas de Monte Carlo para resolver problemas de roteamento de veículos reais é limitada pelo tempo de computação necessário. Por se basear em Simulações de Monte Carlo, diversas soluções são geradas pelo algoritmo para se encontrar soluções ótimas ou próximas de ótimas. Nos testes, foi estipulado um limite de 5 minutos que se mostrou suficiente para encontrar boas soluções, mas que pode impossibilitar a aplicação dessas técnicas em ambientes onde há restrições de tempo.

Podemos notar, entretanto, o desempenho superior dos dois algoritmos que utilizam essas técnicas, o *BinaryMCS-CWS* e o *Monte Carlo Savings*, esse último proposto neste trabalho.

Os dois algoritmos utilizam o *CWS* como base. Enquanto o *BinaryMCS-CWS* explora a árvore de busca fazendo simulações com a mesma lista de *savings*, o algoritmo *Monte Carlo Savings* utiliza as simulações para gerar uma população de lista de *savings* diferentes, se baseando na eficiência bem estudada do algoritmo de *Clarke & Wright Savings* na geração da lista e obtendo, no geral, resultados melhores que todos os outros algoritmos analisados.

8. CONCLUSÃO

Neste trabalho, foi proposto, implementado e analisado o algoritmo *Monte Carlo Savings*, que utiliza uma abordagem simples para aplicação dos Métodos de Monte Carlo, mais especificamente a Simulação de Monte Carlo, em problemas de roteamento de veículos capacitados, baseado no algoritmo heurístico *Clarke & Wright Savings*.

Nos dois *benchmarks* usados, o algoritmo proposto obteve o melhor desempenho entre todos os algoritmos comparados, com variação de apenas 1,39% e 3,13% para as soluções ótimas de cada um dos *benchmarks*, superando métodos bem conhecidos, como a *Busca Tabu Granular*. Em comparação com outros algoritmos que utilizam Simulações de Monte Carlo, o *Monte Carlo Savings* também se mostrou uma boa opção, superando em ambos os *benchmarks* o algoritmo *BinaryMCS-CWS*. Portanto, o algoritmo proposto demonstrou resultados comparáveis aos melhores algoritmos disponíveis na literatura, podendo ser considerado o *estado-da-arte* quando o assunto é Métodos de Monte Carlo para

CVRP.

Trabalhos futuros podem envolver a utilização do algoritmo proposto como entrada para outros métodos meta-heurísticos, incluindo a *Busca Tabu Granular* e o *BinaryMCS-CWS*; estudos e variações da fórmula de cálculo de *savings*, assim como o estudo da sensibilidade da solução proposta em função dos novos parâmetros adicionados, r e λ .

9. REFERÊNCIAS

- [1] TSPLIB, 2014. Disponível em <<http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/>>. Acesso em: 24 ago 2014.
- [2] VRPLIB: A Vehicle Routing Problem LIBrary, 2014. Disponível em <http://www.or.deis.unibo.it/research_pages/ORinstances/VRPLIB/VRPLIB.html>. Acesso em: 24 ago 2014.
- [3] P. Augerat. *Approche polyédrale du problème de tournées de véhicules*. PhD thesis, Institut National Polytechnique de Grenoble-INPG, 1995. Disponível em <<http://www.branchandcut.org/VRP/data>>. Acesso em: 24 ago 2014.
- [4] B. M. Baker and M. Ayeche. A genetic algorithm for the vehicle routing problem. *Computers & Operations Research*, 30(5):787–800, 2003.
- [5] V. P. Barbosa and R. M. Takakura. Implementação e Comparação de Algoritmos que Resolvam o Problema do Roteamento de Veículos Capacitados, 2014. Bachelor's thesis, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, Brasil. 44 f.
- [6] J. E. Bell and P. R. McMullen. Ant colony optimization techniques for the vehicle routing problem. *Advanced Engineering Informatics*, 18(1):41–48, 2004.
- [7] D. Bindel and J. Goodman. *Principles of Scientific Computing*. Manuscript, 2009.
- [8] W. d. O. Bussab and P. A. Morettin. *Estatística básica*. Saraiva, São Paulo, 5 edition, 2004.
- [9] J. S. Christie, S. Satir, and T. P. Campus. Saving our energy sources and meeting Kyoto emission reduction targets while minimizing costs with application of vehicle logistics optimization. In *2006 Annual conference and exhibition of the Transportation Association of Canada: Transportation without boundaries*, Charlottetown, Prince Edward Island, 2006.
- [10] G. Clarke and J. W. Wright. Scheduling of vehicles from a central depot to a number of delivery points. *Operations research*, 12(4):568–581, 1964.
- [11] T. A. Cruse. *Reliability-based mechanical design*, volume 108. CRC Press, 1997.
- [12] G. B. Dantzig and J. H. Ramser. The Truck Dispatching Problem. *Management Science*, 6(1):80–91, 1959.
- [13] J. Faulin and A. Juan. ALGACEA-2: An entropy-based heuristics for the Capacitated Vehicle Routing Problem. In *Seventh Metaheuristics International Conference*, 2007. 3 p.
- [14] M. Fischetti, P. Toth, and D. Vigo. A branch-and-bound algorithm for the capacitated

- vehicle routing problem on directed graphs. *Operations Research*, 42(5):846–859, 1994.
- [15] J. E. Gentle. *Random number generation and Monte Carlo methods*. Springer Science & Business Media, 2003.
- [16] G. M. Giaglis et al. Minimizing logistics risk through real-time vehicle routing and mobile technologies: Research to date and future trends. *International Journal of Physical Distribution & Logistics Management*, 34(9):749–764, 2004.
- [17] G. Gutin and A. P. Punnen, editors. *The traveling salesman problem and its variations*. Combinatorial optimization. Kluwer Academic, Dordrecht, London, 2002.
- [18] J. Y. Kanda. Sistema de meta-aprendizado para a seleção de meta-heurística para o problema do caixeiro viajante. In *Anais do X Simpósio Brasileiro de Sistemas de Informação (SBSI), Londrina*, pages 651–662, 2014.
- [19] G. Laporte. The vehicle routing problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, 59(3):345–358, 1992.
- [20] J. Leeuwen. *Handbook of theoretical computer science: Algorithms and complexity*, volume 1. Elsevier, 1990.
- [21] J. K. Lenstra and A. H. G. Rinnooy Kan. Complexity of Vehicle Routing and Scheduling Problems. *Networks*, 11:221–227, 1981.
- [22] J. Lygaard, A. N. Letchford, and R. W. Eglese. A new branch-and-cut algorithm for the capacitated vehicle routing problem. *Mathematical Programming*, 100(2):423–445, 2004.
- [23] N. Metropolis and S. Ulam. The Monte Carlo Method. *Journal of the American statistical association*, 44(247):335–341, 1949.
- [24] C. Z. Mooney. *Monte Carlo Simulation*. Number 116. Sage, 1997.
- [25] I. H. Osman and J. P. Kelly. Meta-heuristics: An overview. In *Meta-Heuristics*, pages 1–21. Springer, 1996.
- [26] J. Pearl. *Heuristics: Intelligent search strategies for computer problem solving*. Addison-Wesley Pub. Co., Inc., Reading, MA, 1984.
- [27] G. Perboli, R. Tadei, and D. Vigo. The two-echelon capacitated vehicle routing problem: Models and math-based heuristics. *Transportation Science*, 45(3):364–380, 2011.
- [28] K. Shin and S. Han. A Centroid-based Heuristic Algorithm for the Capacitated Vehicle Routing Problem. *Computing and Informatics*, 30(4):721–732, 2011.
- [29] F. Takes and W. Kusters. Applying Monte Carlo Techniques to the Capacitated Vehicle Routing Problem. Master’s thesis, Leiden University, Leiden, Holanda, 2010. 61 f.
- [30] F. Takes and W. Kusters. Applying Monte Carlo techniques to the Capacitated Vehicle Routing Problem. In *Proceedings of 22th Benelux Conference on Artificial Intelligence (BNAIC 2010)*, 2010. 8 p.
- [31] I. J. Taneja et al. On generalized information and divergence measures and their applications: A brief review. *Questiúo: Quaderns d’Estadística, Sistemes, Informàtica i Investigació Operativa*, 13(1):47–73, 1989.
- [32] P. Toth and D. Vigo. *The Vehicle Routing Problem*, pages 1–26. 9 edition, 2002.
- [33] P. Toth and D. Vigo. The granular tabu search and its application to the vehicle-routing problem. *Informa Journal on computing*, 15(4):333–346, 2003.

Recognition of Compromised Accounts on Twitter

Rodrigo Augusto Igawa
Londrina State University
PR 445 Km 380
Londrina-PR, Brazil
igawa.rodrigo@gmail.com

Bruno Bogaz Zarpelão
Londrina State University
PR 445 Km 380
Londrina-PR, Brazil
brunozarpelao@gmail.com

Alex Marino Gonçalves de Almeida
Londrina State University
PR 445 Km 380
Londrina-PR, Brazil
alex.marino.almeida@gmail.com

Sylvio Barbon Jr
Londrina State University
PR 445 Km 380
Londrina-PR, Brazil
barbon@uel.br

ABSTRACT

In this work, we propose an approach for recognition of compromised Twitter accounts based on Authorship Verification. Our solution can detect accounts that became compromised by analysing their user writing styles. This way, when an account content does not match its user writing style, we affirm that the account has been compromised, similar to Authorship Verification. Our approach follows the profile-based paradigm and uses N-grams as its kernel. Then, a threshold is found to represent the boundary of an account writing style. Experiments were performed using a subsampled dataset from Twitter. Experimental results showed that the developed model is very suitable for compromised recognition of Online Social Networks accounts due to the capability of recognize user styles over 95% accuracy.

Categories and Subject Descriptors

I.2.7 [Natural Language Processing]: Text analysis; I.5.4 [Applications]: Text processing; K.4.2 [Social Issues]: Abuse and crime involving computers

General Terms

Measurement, Security, Verification

Keywords

Authoship Verification, Compromised Accounts, N-grams

1. INTRODUCTION

Online Social Networks (OSNs) are environments where people discuss and express thoughts and opinions about any subject [26]. Currently, OSNs represent a relevant resource of information and researches in areas such as Customer Relationship Management (CRM) and Opinion Mining (OM).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

Knowledge obtained from OSNs such as Twitter and Facebook has shown to be extremely valuable for marketing research companies, public opinion organisations, and other Text Mining purposes [1, 25, 28, 19]. Since millions of opinions on a certain topic are expressed with simplicity, posting provide rich, easy and unbiased content comprehension [9]. Therefore, the contents of OSNs are a valuable dataset for decision making on marketing research, business intelligence, stock market prediction and image monitoring [15, 10].

The OSNs wide popularity and ease of access have resulted in the misuse of their services. In addition to the privacy preserving issues, OSNs face the challenge of dealing with undesirable users and their malicious activities, spamming for product promotion being one of the most common [2]. To address the problem of malicious activity on social networks, researchers have focused the detection of fake accounts (i.e., automatically created accounts for only spreading malicious content). However, the problem persists once systems that solely detect fake accounts do not discriminate between fake and compromised accounts. A compromised account is a legitimate account which has been taken over by an attacker to publish fake and harmful content^{1 2}. Accounts can be compromised in many different ways, for example, by exploiting a cross-site scripting vulnerability or by using a phishing scam to steal the users credentials. Also, bots have been increasingly used to obtain credentials information for social networking sites on infected hosts [6, 8].

Since fake accounts were mainly created with proposal to cause harm in OSNs, once they are detected, the simplest solution is to delete them. In the meantime, compromised accounts need engaging in a credentials recovery process to give back the accounts control to their respective owners [6]. Moreover, a study performed through Twitter revealed that only 16% of the spamming accounts were indeed fake accounts, while the remaining quantity were all compromised accounts [8]. The same reality also was seen on Facebook where 97% of malicious accounts were not originally created to solely spamming purpose [7].

In this paper, we present a first of its kind study to recognize compromised accounts using just text as resource. Our approach is based on N-grams Authorship Verification (AV) and we focus on recognition of a user based on its writing

¹<http://www.bbc.com/news/world-us-canada-30853311>

²<http://www.bbc.com/news/world-us-canada-30785232>

style. When the writing style of a given user does not match its boundary based on a threshold, then, a warning alarm could be sent out to inform the account owner and malicious posts could be blocked. Also, as seen in [13, 23, 15] Preprocessing can either contribute or disturb text mining tasks, therefore, we also conducted experiments concerning Preprocessing and Corpus size to study their relevance in results. Our experiments were performed using a Twitter dataset and results ranging from 94% to 95% accuracy were achieved.

The remaining of the work is organized as follows: Section 2 presents an overview towards compromised accounts and Authorship Verification along N-grams. In Section 3 details about the proposed approach are described. Section 4 presents the experimental settings to perform our tests. Section 5 discusses our results. Section 6 states our conclusions.

2. RELATED WORK

Compromised accounts initially became the object of research interest in e-mail and web services as seen in [22, 12]. In a similar scenario to OSNs, users credentials are stolen using malicious link or phishing techniques [14, 22]. Concerning e-mails, researches already conducted work in user levels by using social engineering to emphasize user awareness [12], while another different approach combined network information, machine learning and content analysis in order to detect harmful content [22].

Some other approaches detected intrusion and compromised accounts in short messages by applying text mining techniques as Authorship Attribution (AA) and AV [5, 4]. Their main contribution was to aid the search for cyber criminals [27] or to increase cyber space security and reliability [5].

In order to achieve so, both AA and AV were based on one of two strategies: Stylometry and N-grams. The first one describes text content through attributes which represent writing-style markers as lexical, syntactic, content-specific, and idiosyncratic style markers. Lexical attributes are words and character based statistical measures like sentence length. Syntactic attributes include part-of-speech tagger measures. Content-specific attributes are represented by keywords of a given text and idiosyncratic markers are represented by misspellings and grammatical mistakes [11, 18].

N-grams, on the other hand, consist in obtaining frequent co-occurrent patterns in words or character level. A set of most frequent N-grams represents the textual description of a given author, hoping that most frequent patterns would occur more often [13, 21].

Regarding the few existent works addressing compromised accounts on OSNs, studies already stated that malicious content are almost completely spread by compromised accounts that were victims of phishing attacks. The detection of malicious accounts is achieved by extracting features from text, webdata and network information to then, classify it based on machine learning approaches like Random Forest, SVM and Logistic Regression [7, 20].

3. PROPOSED APPROACH

The proposed approach is grounded on AV to analyse if an account has been compromised. In order to represent the legitimate user, it is necessary to extract features from tex-

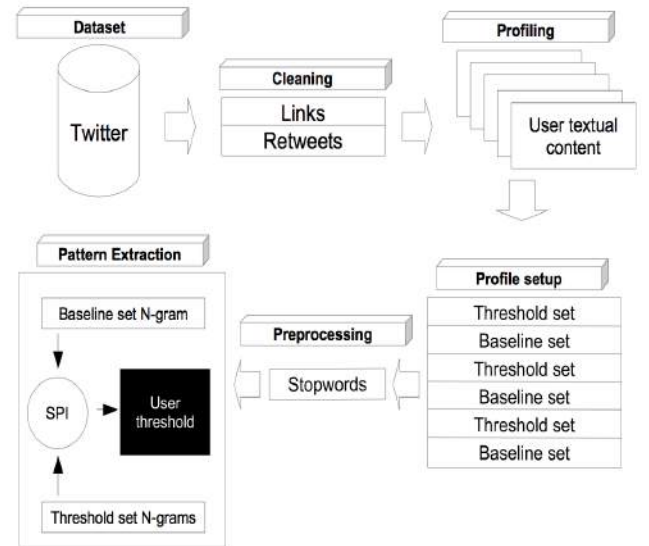


Figure 1: Proposed Approach for User Threshold Estimation

tual content. These contents are calculated using N-grams, as seen in in Figure 1.

The main idea behind our proposal is to address compromised accounts problem as a document representation model. By doing so, it would be possible to apply Text Mining tasks to analyze the user writing style.

First step is dataset acquisition. Considering Text Mining approach only text features will be used and, therefore, none additional information beyond the tweets content and their respective authors username are required for our proposal.

Second step is about Cleaning. Normally, it would be possible to consider any textual content as a part of a document produced by an author. However, as this approach is created to be applied on Twitter users, links and *retweets* (thirdy part contents) were removed due to not represent any textual authorship mark.

All remaining text productions are considered authorship samples. Therefore, all contents are concatenated cumulatively following the profile-based paradigm described in [17]. The result of this step, called Profiling, is a document containing all terms written by the user.

Then, in Profile Setup, each user is represented as a document whose content is subsampled at the same fraction in two distinct parts: Baseline set and Thresholding set. Each fraction of document subsampled has the same size and is interspersed as shown in Figure 1. This way, subjects discussed by the user during the time will be equally distributed in both sets. This is important to our approach because both sets must have subjects balanced so that the boundary of the writing style can be properly found.

The Baseline set is the text portion which represents user account. This set is used to extract the usual writing style of an user and is kept as one single document as described by the profile-based paradigm in [17]. The Thresholding set is a portion used to find a Simplified Profile Intersection (SPI) threshold to delimitate the user writing style and different from Baseline set, each portion subsampled becomes an distinct sample instance. The SPI similarity measure was used in [17, 13] and is stated to be suitable to diferent sample

sizes. The SPI is calculated as seen in Equation 1, where N_1 and N_2 are two distinct sets of N-grams. Note that SPI is basically a count of N-grams that exist in both sets.

$$SPI(N_1, N_2) = |N_1 \cap N_2| \quad (1)$$

After Profile Setup process, Preprocessing techniques can be performed to improve the effectiveness of our approach. In this work, we explore some combination of Preprocessing concerning precision and accuracy to recognize accounts textual content.

In order to obtain the SPI threshold in Writing Style Extracion step, most frequents N-grams are extracted from Baseline set and most frequents N-grams are also extracted from each fraction in Thresholding set. Then, SPI is used to calculate similarity between Baseline set and Thresholding set N-grams. The minimal similarity obtained is considered the SPI threshold. Any future portion of text posted in this account that presents similarity measure lesser than threshold is considered an intrusion and the account is compromised.

4. METHODOLOGY

Twitter, the OSN used in this work, is known as a micro blogging service. Unlike other social media, Twitter is known by short posts (140 characters at maximum) done by users expressing thoughts, opinions and feelings [26]. These short texts, named tweets, are available publicly as default, and are immediately broadcasted to the users followers [3].

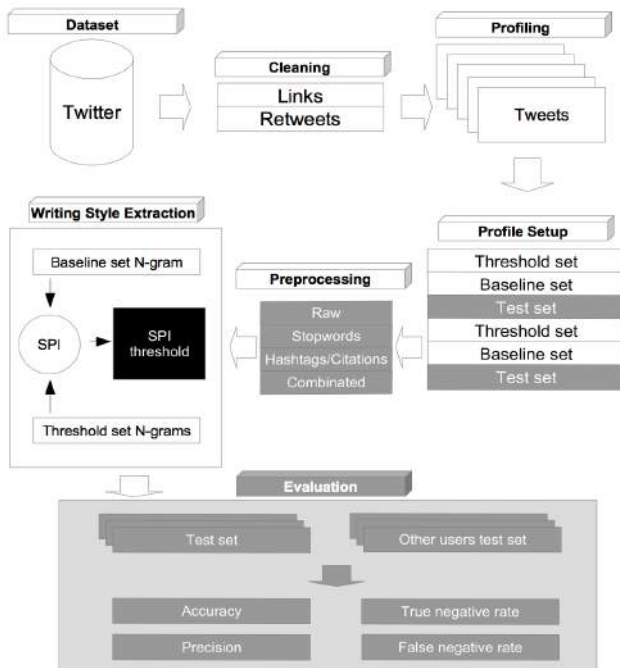


Figure 2: Experimental Settings Overview

The Twitter Developer Team offer a streaming service that delivers other developers low latency access to Twitter's global stream of data. The tweets sets used as samples in our experiments were collected by [24] using this service.

In our experiments, only a subsample from original Dataset was used. This was done in order to simulate the few

samples available in a more critical scenario. From this part, all tweets were grouped by authors username cumulatively following profiling step described in Section 3. As specified in our approach, links and retweets were removed since they do not infer any information about its user writing pattern. All remaining textual content was included in our tests.

In Figure 2, gray parts represent the experimental settings. The Profile Setup process was performed to separate textual contents in 3 distincts parts: Baseline set, Thresholding set (as proposed in Section 3) and Test set. The last set is used to evaluate our method's efficiency and is a representation of future portions of text posted. The user's Test set is completely used along randomly selected Test set instances from other users to check how adequate is the obtained threshold. Our intention to use other users test set among the own user test set is to simulate a situation where the legitimate account has been compromised and harmful posts are written, therefore, it is desirable to obtain SPI measures in instances from other users test set lesser than threshold, while own user test set are intended to present SPI greater than threshold.

Another concerning towards the Profile setup is to study the size of each splitted part. This is considered an important issue of this work once the size used presenting better results would be the amount of words necessary to recognize accounts textual contents. In our experiments, were used 11 difrent sizes ranging from 50 to 100 words.

Also, concerning Preprocessing, 4 tests concerning their influence were performed: a) Raw (i.e, no preprocessing), b) *Hashtags* and *Citations* removal, c) Stopwords removal and d) Combinated preprocessing (i.e, *Hashtags*, *Citations* and Stopwords removal). The idea behind these tests is to study the influence of disposable terms concerning precision and accuracy to recognize account textual content.

One last issue experienced in our tests was the N used on N-grams. Following results from [13] we used values to N equals to 4, 5 and 6. These setting were also applied including the Corpus size and preprocessing settings.

Therefore, the complete experimental setting consists in 132 experiments covering our 3 diferent N-grams values, 11 combination towards Corpus size the 4 combination dealing 2 preprocessing techniques.

In order to keep experiments always balanced to enable comparisons to each other, we defined that each Test set and Threshold set were composed by 10 instances of same size (ranging from 50 to 100 words). In the Evaluation step, the user being recognized always used its entire Test set (i.e, 10 instances of text) along 10 instances randomly selected from other users Test sets.

To evaluate our method efficiency, we used 4 well known statistical measures found in [16] and their equations are shown in Table 1 where TP are user test set instances presenting SPI greater than threshold, TN are other users test set presenting SPI lesser than threshold, FN are user test instances presenting SPI lesser than threshold and FP are other users test instances presenting SPI greater than threshold. Analysis results and discussion towards all experiments are presented in Section 5.

5. RESULTS AND DISCUSSION

As described previously, 132 experiments were performed concerning all possible combinations within $N = 4, 5, 6$; Corpus size in each splitted portion ranged from 50 to 100 words

Table 1: Measures used to evaluate recognition rate

Name	Equation
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$
False Negative Rate	$\frac{FN}{FN+TP}$
True Negative Rate	$\frac{TN}{TN+FP}$

Table 3: The influence of text Preprocessing techniques on compromised accounts recognition

Preprocessing	Mean	Standard Deviation
Raw	91.90%	7.21%
Hashtags/Citations Removal	86.10%	10.56%
Stopwords Removal	95.80%	8.78%
Combinated Preproc.	85.22%	12.15%

and 4 combinations of text preprocessing. Table 2a and 2b shows our highest and lowest results in accuracy terms independently of its setting. It is notable that the top settings achieved excellent results, ranging from 94.10% to 95.80% accuracy (i.e. correctly classified instances) and also present excellent results in terms of true negative rate ranging from 88.40% to 91.60% which indicates that our method is capable of infer when the content does not correspond to its legitimate user writing pattern.

Another important issue to be observed is the combinations of preprocessing in both Table 2a and Table 2b. All 10 top results achieved their results without removing hashtags and citations. On the other hand, all 10 least accurate experiments applied hashtags and citations removal achieving poor results. Our first conclusion by overviewing the experiments is that hashtags and citations carry information about the writing style of an user textual content, once they indicate subjects discussed and people frequently contacted.

Still concerning the preprocessing issue, a detailed result from the top 1 experimental setting in Table 2a using Corpus size = 100 and $N = 6$ is shown in Table 3 in terms of accuracy. Just by removing hashtags and citations, a loss in accuracy is found, falling from 91.90% to 86.10% precision. By removing only stopwords it is still possible to increase 5.0% accuracy. This implies that pronouns, articles and prepositions used do not help to recognize an user writing style using our approach. One last observation about preprocessing is: a combination of hashtags/citations and stopwords removal achieves the lowest results of the 4 combinations once it uses only a little part of writing not including stopwords, hashtags and citations.

A discussion towards the top 1 setting in Table 2a is illustrated by Figure 3 and shows accuracy considering each user. The setting achieved 100% of correctly recognized in many cases, however, to account number 4, 15, 25 and 27 obtained accuracy below 80%. These users presented a very unstable writing style using a high quantity of prepositions and almost nothing of jargons and emoticons making their writing difficult to distinguish. In all other cases the setting obtained satisfactory results.

Corpus size influence on our approach is illustrated by Figure 4. Before any discussion about this view, it is neces-

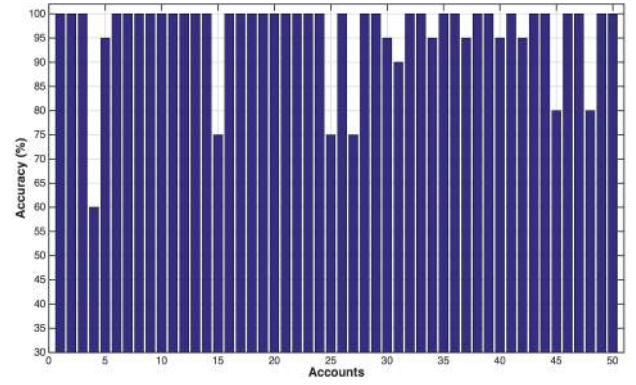


Figure 3: Individual accuracy on recognition of compromised accounts

sary to observe that the Corpus size is not used only to split in Profile Setup process, but also implies in the number of words necessary to perform proposed approach with satisfactory results. Considering so, the fact that none setting size used in our experiments presented outliers and also presented balanced quartiles, having a good result. It implies that our method have a stable range of accuracy independently of amount of text used. A descending gradient observed on accuracy using 100 to 50 words is justifiable once less words also means less n-grams to be extracted and possibly less accuracy. Therefore, the boxplot states that the most considerable size to be used in our dataset is 100 words while the most inappropriate is 50.

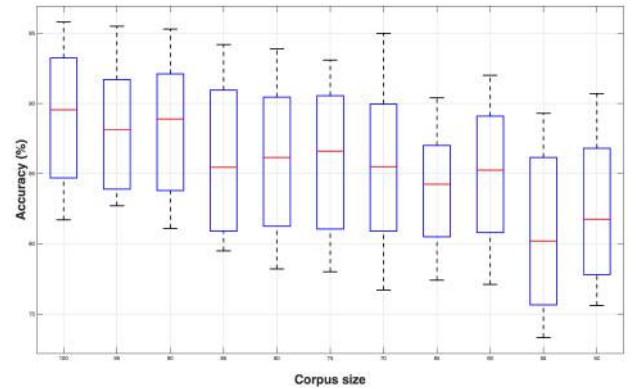


Figure 4: The influence of Corpus size on baseline accuracy

One last consideration towards our approach is threshold value and its relation to each user writing style. Most part of users obtained 100% accuracy as shown in Figure 3. These users are represented in Figure 5, as Case III, where the obtained threshold is suitable to separate writing styles from the legitimate user and other users. Case I and II represent users that by presenting too many stopwords as part of their writing styles and using a small quantity of emoticons, jargons, hashtags or citations obtained a threshold value unable to correctly separate writing styles and, therefore, obtained a significant number of false negative (i.e. writing style from different users being recognized as the user in question).

6. CONCLUSION AND FUTURE WORK

Table 2: Overview on accounts recognition accuracy

(a) Top results in accuracy

N	C. Size	Precision	Accuracy	TNR	FNR	Hashtags/Citations	Stopwords
6	100	93.97%	95.80%	91.60%	0.00%	Not removed	Removed
5	100	93.43%	95.70%	91.40%	0.00%	Not removed	Removed
6	95	93.59%	95.50%	91.00%	0.00%	Not removed	Removed
6	90	93.64%	95.30%	90.80%	0.20%	Not removed	Removed
5	90	93.36%	95.10%	90.40%	0.02%	Not removed	Removed
6	70	92.72%	95.00%	90.20%	0.02%	Not removed	Removed
4	100	92.28%	94.60%	89.20%	0.00%	Not removed	Removed
5	70	92.45%	94.60%	89.40%	0.02%	Not removed	Removed
6	85	92.57%	94.20%	89.00%	0.06%	Not removed	Removed
5	95	91.39%	94.10%	88.40%	0.02%	Not removed	Removed

(b) Lowest results in accuracy

N	C. Size	Precision	Accuracy	TNR	FNR	Hashtags/Citations	Stopwords
6	55	72.31%	77.60%	55.40%	0.20%	Removed	Removed
6	65	73.75%	77.40%	61.20%	6.40%	Removed	Not removed
6	50	75.08%	77.20%	63.20%	8.80%	Removed	Not removed
4	60	75.47%	77.10%	64.40%	10.20%	Removed	Not removed
4	70	74.53%	76.70%	61.80%	8.40%	Removed	Not removed
4	55	71.73%	76.60%	53.80%	0.60%	Removed	Removed
5	50	74.03%	76.10%	60.60%	8.40%	Removed	Not removed
4	50	73.89%	75.60%	60.20%	9.00%	Removed	Not removed
4	55	71.30%	74.70%	55.80%	6.40%	Removed	Not removed
6	55	70.97%	73.80%	55.40%	7.80%	Removed	Not removed
5	55	70.62%	73.30%	55.00%	8.40%	Removed	Not removed

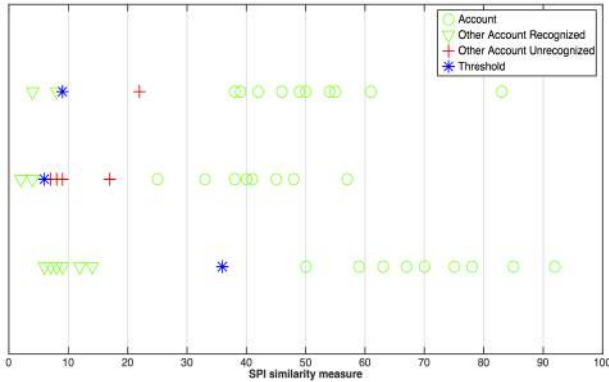


Figure 5: Threshold testing for compromised accounts recognition

One advantage from our approach is that only text is used as resource once it is grounded on Text Mining. Although it was tested on Twitter in our experiments, our developed method is applicable in any OSN. Also, due to the fact that this work is the first to depend only on text to recognize compromised accounts, our approach concerned about the Corpus size necessary to recognize compromised accounts, desirable preprocessing to obtain better results and which N use in N-grams calculation to improve the approach results.

In this work one important consideration during the entire process is that warning systems should not incorrectly

recognize a legitimate user as an invader to claim compromised account. Therefore, we studied and comproved that the top experimental setting presented in Table 2a (N=6, Corpus size = 100, Stopwords removal only) obtains very good results in terms not only of precision and accuracy, but also very few occurrences of false negative. That implies that our method would rarely claim compromised accounts by content when actually it was not compromised.

Considering Twitter's very short texts scenario, the need to use 100 words is also acceptable. The tweets used in experiments have a mean of 14.6 words. In practice, between 6-10 tweets, depending of how much per tweets is written, is possible to recognize an user by its text, with 95% accuracy along 91% true negatives.

For future works it would be of great interest to study a method dealing only with those cases of low accuracy presentend in Figure 5 and Figure 3. This way, our methods accuracy could be increase. This study could be towards other N-grams measures focused on special cases of authors.

7. REFERENCES

- [1] S.-A. Bahrainian and A. Dengel. Sentiment analysis and summarization of twitter data. In *Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on*, pages 227–234. IEEE, 2013.
- [2] S. Y. Bhat and M. Abulaish. Community-based features for identifying spammers in online social networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social*

- Networks Analysis and Mining*, pages 100–107. ACM, 2013.
- [3] C. A. Bliss, I. M. Kloumann, K. D. Harris, C. M. Danforth, and P. S. Dodds. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*, 3(5):388–397, 2012.
- [4] M. L. Brocardo, I. Traore, S. Saad, and I. Woungang. Authorship verification for short messages using stylometry. In *Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on*, pages 1–6. IEEE, 2013.
- [5] J. A. Donais, R. A. Frost, S. M. Peelar, and R. A. Roddy. Summary: A system for the automated author attribution of text and instant messages. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 1484–1485. IEEE, 2013.
- [6] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna. Compa: Detecting compromised accounts on social networks. In *NDSS*, 2013.
- [7] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 35–47. ACM, 2010.
- [8] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37. ACM, 2010.
- [9] A. Hassan, A. Abbasi, and D. Zeng. Twitter sentiment analysis: A bootstrap ensemble framework. In *Social Computing (SocialCom), 2013 International Conference on*, pages 357–364. IEEE, 2013.
- [10] L.-C. Hsieh, C.-W. Lee, T.-H. Chiu, and W. Hsu. Live semantic sport highlight detection based on analyzing tweets of twitter. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 949–954. IEEE, 2012.
- [11] S. Keretna, A. Hossny, and D. Creighton. Recognising user identity in twitter social networks via text mining. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pages 3079–3082. IEEE, 2013.
- [12] S. Khanna and H. Chaudhry. Anatomy of compromising email accounts. In *Information and Automation (ICIA), 2012 International Conference on*, pages 640–645, June 2012.
- [13] R. Layton, P. Watters, and R. Dazeley. Authorship attribution for twitter in 140 characters or less. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second*, pages 1–8. IEEE, 2010.
- [14] C.-H. Li, F.-H. Hsu, S.-J. Chen, C.-S. Wang, Y.-H. Chen, and Y.-L. Hwang. Hawkeye: Finding spamming accounts. In *Network Operations and Management Symposium (APNOMS), 2014 16th Asia-Pacific*, pages 1–4. IEEE, 2014.
- [15] M. M. Mostafa. More than words: Social networks’ text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241–4251, 2013.
- [16] D. L. Olson and D. Delen. *Advanced data mining techniques*. Springer Science & Business Media, 2008.
- [17] N. Potha and E. Stamatatos. A profile-based method for authorship verification. In *Artificial Intelligence: Methods and Applications*, pages 313–326. Springer, 2014.
- [18] R. Ramezani, N. Sheydaei, and M. Kahani. Evaluating the effects of textual features on authorship attribution accuracy. In *Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on*, pages 108–113. IEEE, 2013.
- [19] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič. Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 2014.
- [20] T. Stein, E. Chen, and K. Mangla. Facebook immune system. In *Proceedings of the 4th Workshop on Social Network Systems*, page 8. ACM, 2011.
- [21] J. Sun, Z. Yang, P. Wang, and S. Liu. Variable length character n-gram approach for online writeprint identification. In *Multimedia Information Networking and Security (MINES), 2010 International Conference on*, pages 486–490. IEEE, 2010.
- [22] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and evaluation of a real-time url spam filtering service. In *Security and Privacy (SP), 2011 IEEE Symposium on*, pages 447–462. IEEE, 2011.
- [23] A. K. Uysal and S. Gunal. The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112, 2014.
- [24] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.
- [25] S. J. Yu. The dynamic competitive recommendation algorithm in social network services. *Information Sciences*, 187:1–14, 2012.
- [26] M. Zappavigna. Ambient affiliation: A linguistic perspective on twitter. *New Media & Society*, 13(5):788–806, 2011.
- [27] C. Zhang, X. Wu, Z. Niu, and W. Ding. Authorship identification from unstructured texts. *Knowledge-Based Systems*, 2014.
- [28] X. Zhou, S. Wu, C. Chen, G. Chen, and S. Ying. Real-time recommendation for microblogs. *Information Sciences*, 279:301–325, 2014.

Avaliação de um Modelo de Maturidade para Governança Ágil em TIC usando *Focus Group*

Alternative title: Evaluation of a Maturity Model for Agile Governance in ICT using Focus Group

Humberto Rocha de
Almeida Neto

Centro de Informática - UFPE
CEP: 50740-540 - Recife - PE -
Brasil
hran@cin.ufpe.br

Edviges Mariza Campos de
Magalhães

DATAPREV
CEP: 58040-000 - João Pessoa - PB
- Brasil
edviges.magalhaes@dataprev.gov.br

Hermano Perrelli de
Moura

Centro de Informática - UFPE
CEP: 50740-540 - Recife - PE -
Brasil
hermano@cin.ufpe.br

José Gilson de Almeida
Teixeira Filho

Centro de Informática - UFPE
CEP: 50740-540 - Recife - PE -
Brasil
jgattf@cin.ufpe.br

Claudia Cappelli

Departamento de Informática
Aplicada - UNIRIO
CEP: 22290-240 - Rio de Janeiro -
RJ - Brasil
claudia.cappelli@uniriotec.br

Luiz M. Fraga Martins
Departamento de Engenharia
de Informática - CISUC
Pólo II. 3030-290 - Coimbra -
Portugal
lmart@dei.uc.pt

RESUMO

A Governança Ágil em Tecnologia da Informação e Comunicação (TIC) é uma área em ascensão e vem sendo sugerida como uma proposta inovadora baseada no uso de princípios e valores do Manifesto para Desenvolvimento Ágil de Software sobre o contexto da Governança em TIC convencional. No entanto, a adoção de práticas para uma boa Governança Ágil em TIC ainda é considerada uma tarefa desafiadora devido, principalmente, a multidisciplinaridade e relativa incipiência da área. Se por um lado, há um conjunto de princípios, práticas e valores de áreas subjacentes a este domínio, por outro lado, esse conjunto ainda não se apresenta organizado em um contexto mais amplo de forma a promover um caminho sistemático e gradual para uma evolução de maturidade em Governança Ágil em TIC. Diante desta lacuna, este artigo apresenta uma proposta de modelo de maturidade para Governança Ágil em TIC construído a partir de um amplo estudo bibliográfico envolvendo as áreas de Governança em TIC, agilidade e maturidade. Durante a construção do modelo proposto, partiu-se de uma abordagem qualitativa, com método indutivo e procedimentos comparativo-estruturalistas. Por fim, este artigo apresenta uma avaliação do modelo proposto por meio de *focus group*. Como resultado, um conjunto de melhorias identificadas durante o estudo vem subsidiando a construção de uma nova versão do modelo.

Palavras-Chave

Governança Ágil; Governança em Tecnologia da Informação e Comunicação; Modelos; Maturidade; *Agile*.

ABSTRACT

Agile Governance in the Information and Communication Technology (ICT) is an area on the rise and has been suggested as an innovative proposal based on the use of principles and

values of the Manifesto for Agile Software Development on Governance in ICT conventional context. However, the adoption of practices to good Agile Governance in ICT is still considered a challenging task due mainly multidisciplinary and relative paucity of the area. On the one hand, there is a set of principles, practices and values of the underlying areas this domain, on the other hand, this set is not presented yet organized in a broader context in order to promote a systematic and gradual way to an increase of maturity in Agile Governance in ICT. Given this gap, this article proposes a maturity model for Agile Governance in ICT constructed from a wide bibliographic study involving the areas of Governance in ICT, agility and maturity. During construction of the proposed model, it started with a qualitative approach, with inductive method and comparative-structuralist procedures. Finally, this paper presents an evaluation of the model proposed by focus group. As a result, a number of improvements identified during the study is subsidizing the construction of a new version of the model.

Categories and Subject Descriptors

[**Management**]: IT Governance – *agile governance in information and communication technology*.

General Terms

Management, Measurement, Performance, Reliability, Experimentation, Human Factors, Legal Aspects.

Keywords

Agile Governance; Governance in Information and Communication Technology; Models; Maturity; *Agile*.

1. INTRODUÇÃO

Iniciativas diversas têm sido propostas com objetivo de produzir uma Governança em Tecnologia da Informação e Comunicação (TIC) mais eficiente (GREMBERGEN; DEHAES, 2008; ISACA, 2009; WEIL; ROSS, 2005).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.
Copyright SBC 2015.

De acordo com o IT Governance Institute (ITGI, 2001), a Governança em TIC é definida como um subconjunto da Governança Corporativa, sendo considerada uma disciplina focada em TIC e seus sistemas de desempenho e gestão de riscos. A Governança Corporativa, por sua vez, pode ser encarada como o conjunto de processos, costumes, políticas, leis e instituições que afetam a maneira como uma empresa é dirigida, administrada ou controlada. Isto inclui as relações entre as várias partes envolvidas e os fins para os quais a sociedade é governada (CADBURY, 1992).

Diante deste contexto, boas práticas, modelos, guias e *frameworks* dos mais variados têm sido propostos. Entre estes podemos citar: a ITIL v3 (OGC, 2009) e o COBIT 5 (ISACA, 2012). A biblioteca de boas práticas da ITIL v3 apresenta contribuições para apoio ao Gerenciamento de Serviços de TIC, enquanto que o *business framework* do COBIT 5, foca no apoio ao alinhamento entre a TIC e o negócio e a garantia da integridade de seus sistemas e informações.

Apesar de contribuírem para o apoio a Governança em TIC, alguns autores afirmam que a grande maioria destas iniciativas ainda são burocráticas, consomem muito tempo e não apresentam detalhes de orientação sobre como implementar e implantar os instrumentos de gestão necessários e mecanismos de Governança (KTATA; LÉVESQUE, 2009; LOEWENSTERN et al., 2010).

Neste sentido é possível observar um conflito entre o formalismo apresentado pela grande maioria destas iniciativas e a agilidade imposta por um mercado cada vez mais competitivo. Em meados de 2001, pode-se observar uma dicotomia semelhante no cenário de desenvolvimento de software da época. Naquele período, metodologias como a *Rational Unified Process* (RUP) (RATIONAL, 2004), tidas como precursoras ao desenvolvimento prescritivo de software, também se depararam com um problema semelhante. Este problema motivou o surgimento do *Manifesto for Agile Software Development* (BECK et al., 2001), manifesto que se pautou em um conjunto de princípios e valores inovadores que promoveu uma quebra de paradigmas.

Desta forma, com esta mesma perspectiva em mente, acredita-se que a Governança em TIC também carece atualmente de iniciativas mais ágeis e práticas. LUNA et al., (2010) apresentam um estudo demonstrando a relação entre os fatores críticos de sucesso para Governança em TIC e os princípios do *Manifesto for Agile Software Development*. Por exemplo, para o fator crítico de sucesso: “**F7 - Foco em pequenas-consecutivas vitórias e na disseminação dos resultados destas iniciativas**”, não foram identificadas relações apenas entre 2 dos 12 princípios do *Manifesto for Agile Software Development*. Este estudo reforça a existência de relação entre as áreas de Governança em TIC e agilidade. Maiores detalhes sobre cada um dos princípios ágeis e fatores críticos de sucesso para Governança investigados durante o estudo podem ser obtidos em (LUNA et al., 2010).

Além disto, estes mesmos autores propõem, a partir deste mesmo estudo, uma das definições envolvendo Governança Ágil, uma proposta inovadora, pautada no uso dos princípios e valores do *Manifesto for Agile Software Development* sobre a Governança convencional. De acordo com recentes estudos deste domínio, apesar de promissora, esta pode ser considerada uma área ainda em ascensão, requerendo de iniciativas para o avanço deste campo multidisciplinar e ainda incipiente (LUNA et al., 2014).

Outro fator interessante motivou o desenvolvimento desta proposta de trabalho foi que, durante o desenvolvimento de nossas pesquisas, foi constatado que as metodologias de desenvolvimento de software (ágeis ou convencionais) têm recebido apoio de modelos de maturidade e que o mesmo também tem acontecido com algumas das iniciativas de apoio a Governança em TIC. No entanto, para Governança Ágil em TIC, até o presente momento, não foram identificadas propostas de modelos de maturidade como o proposto nesta pesquisa.

Modelos de maturidade buscam uma unificação de uma mesma visão, tratando a evolução da maturidade como estágios evolutivos. Nestes estágios, as organizações vão evoluindo e conquistando um maior grau de maturidade a cada etapa avançada (PRADO, 2008; SOLER, 2005). A maturidade é definida como a evolução da capacidade do processo, onde cada nível de maturidade provê uma camada para a melhoria contínua do processo (SEI, 2010).

Segundo Carmel (2005), o entendimento coletivo de um fenômeno seria incompleto se os estágios de maturidade não fossem identificados e isto se torna ainda mais relevante quando o fenômeno está nos seus estágios iniciais de estudo: cenário atual do domínio de Governança Ágil em TIC.

Diante deste contexto, este artigo apresenta uma proposta de modelo de maturidade para Governança Ágil em TIC. Acreditamos que a concepção de um modelo de maturidade como que este, traga contribuições de considerável inovação e relevância. Além disto, este trabalho apresenta os resultados de uma das avaliações do modelo proposto utilizando *focus group*. As seções seguintes são estruturadas da seguinte maneira: a seção 2 apresenta a metodologia que vem sendo empregada neste trabalho de pesquisa, a seção 3 apresenta uma visão geral do modelo proposto e a seção 4 apresenta os resultados obtidos em uma de suas avaliações. Por fim, a seção 5 apresenta as devidas conclusões e trabalhos futuros para este trabalho de pesquisa.

2. METODOLOGIA

Marconi e Lakatos (MARCONI; LAKATOS, 2003) afirmam que a definição instrumental metodológica deve estar diretamente relacionada com o problema a ser estudado.

Desta forma, uma pesquisa deve ser rigorosamente analisada antes mesmo de sua execução propriamente dita. Com o intuito de atender ao objetivo principal deste trabalho, foram definidos alguns procedimentos e técnicas conforme **Tabela 1**.

Tabela 1. Resumo das Abordagens Metodológicas.

Objetivo	Exploratória
Procedimentos Técnicos	Pesquisa Bibliográfica; Revisão Sistemática da Literatura; <i>Survey</i> ; <i>Focus Group</i> .
Natureza das Variáveis	Quantitativa; Qualitativa.
Método de Abordagem	Indutiva.
Métodos de Procedimento	Comparativo; Estruturalista.
Áreas de Concentração	Ciências da Computação (Sistemas de Informação); Ciências da Administração (Sistemas de Informação).

2.1 Processo de Pesquisa

Esta pesquisa foi realizada através de um conjunto de atividades estruturadas em duas fases distintas conforme **Figura 1**. A primeira fase é composta por um conjunto de quatro etapas. Sua característica é mais exploratória e tem o objetivo de construção

de um arcabouço teórico consistente para suporte as etapas posteriores da pesquisa.

Em um primeiro momento, buscando obter uma visão geral a respeito dos objetos de investigação, foi realizada uma pesquisa bibliográfica inicial abordando principalmente as áreas centrais da pesquisa. Esta etapa gerou um referencial teórico envolvendo principalmente: Governança em TIC, agilidade no desenvolvimento de software e modelos de maturidade. Em seguida, foram executados dois estudos exploratórios. O primeiro estudo buscou identificar iniciativas para avaliação de maturidade no desenvolvimento de software (ALMEIDA NETO et al., 2012a). O segundo buscou identificar iniciativas para avaliação de maturidade na Governança em TIC (ALMEIDA NETO et al., 2012b, 2012c).

Em seguida, com o objetivo de reforçar a revisão de literatura, foi executada uma Revisão Sistemática da Literatura (RSL). Esta RSL resultou em um levantamento do estado da arte e evoluções da Governança Ágil no mundo (LUNA et al., 2014).

Os resultados obtidos na primeira fase da pesquisa serviram de alicerce para o desenvolvimento das etapas posteriores. A segunda deste trabalho possui uma característica mais descritiva visando à construção e melhoria contínua do modelo proposto. Em uma primeira etapa, foi idealizado um esboço inicial do modelo, sendo este composto por uma arquitetura e seus respectivos componentes. Em seguida, este esboço inicial foi submetido a discussões internas ao nosso grupo de pesquisa e a discussões entre membros das comunidades científicas e da indústria (ALMEIDA NETO; MOURA; CAMPOS DE MAGALHAES, 2014; ALMEIDA NETO; MOURA, 2014). Como fruto destes *feedbacks*, o modelo foi então evoluído para uma versão alfa. Na etapa seguinte, foi executado um *survey* com especialistas, onde a primeira versão do modelo proposto pode ser avaliada por um universo mais amplo (ALMEIDA NETO et al., 2015). As informações coletadas a partir deste estudo serviram de base para geração de uma versão beta do modelo proposto. Por fim, foi então executada uma avaliação por *focus group*. Este estudo gerou contribuições valiosas, a partir de opiniões diversas de profissionais diretamente ligados ao mercado. Atualmente os dados coletados a partir deste estudo estão sendo utilizados para o aprimoramento do modelo proposto.

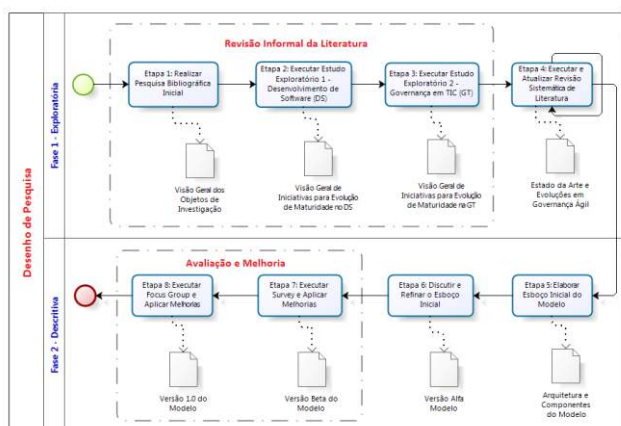


Figura 1. Processo de Pesquisa para construção do modelo.

3. MODELO PROPOSTO

O modelo proposto foi denominado de *MAnGve Maturity Model* (M3) sendo estruturado por meio de três componentes básicos: Um Método de Avaliação (MA), um Modelo de Referência (MR) e uma Base de Experiências (BE). O principal objetivo do MR é a descrição dos processos especificados para o modelo. O MA, por sua vez, é um método que orienta a execução de avaliações de conformidade segundo o MR. Por fim, a BE é tida como um repositório onde são registradas as experiências das avaliações executadas. Além disto, o M3 foi construído para permitir a integração com outras iniciativas para Governança Ágil desenvolvidas no âmbito do mesmo grupo de pesquisa¹ (LUNA, 2009; LUNA, A.J.H. O., KRUCHTEN, P., MOURA, 2013; PEDROSA, 2014).

3.1 Metamodelo do MR

Um modelo conceitual é normalmente utilizado no projeto de sistemas de informação e pretende de uma maneira abstrata demonstrar relacionamentos entre os seus componentes (HEUSER, 1998). No escopo deste trabalho, foi criado um modelo conceitual sob forma de um metamodelo para o MR. Este metamodelo define os elementos básicos que estruturam o MR e o relacionamento entre estes (Figura 2).

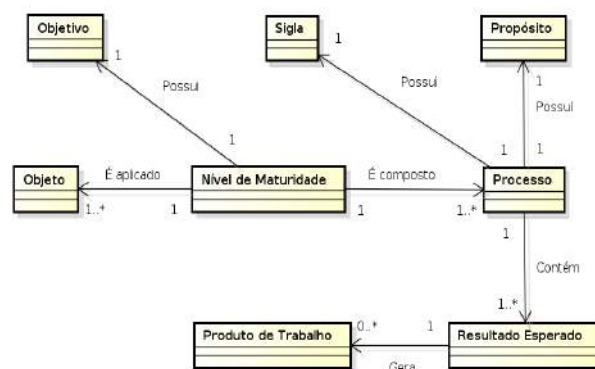


Figura 2. Metamodelo do MR.

Neste metamodelo é possível observar que:

- Cada **Nível de Maturidade** possui um único objetivo, é composto de um ou mais processos e é aplicado a um ou mais objetos.
- O **Objetivo** descreve o propósito do nível de maturidade e está relacionado a um único nível de maturidade.
- O **Objeto** descreve o contexto ao qual pode ser aplicado o nível de maturidade, e deve ser aplicado, no mínimo, a um nível de maturidade. Estes objetos podem ser, por exemplo, a organização como um todo ou mesmo parte dela (um setor, uma diretoria, uma unidade de negócio, entre outras).
- O **Processo** deve conter um conjunto de resultados esperados e um propósito único. Um processo apresenta uma descrição de atuação a partir de um nome de maneira simples e objetiva. Desta forma, um processo pode ser visto como um conjunto de ações correlacionadas que são executadas com

¹ <http://gp2.cin.ufpe.br/>

vistas a gerar resultados esperados. Um processo também possui uma sigla única.

- O **Propósito** está associado a um único processo e direciona a organização para a evolução através da descrição de objetivos que norteiam a realização dos processos.
- Os **Resultados Esperados** definem os resultados a serem obtidos após a execução do processo e podem gerar um conjunto de produtos de trabalho.

Além disto, um ciclo de vida para utilização deste modelo de maturidade também foi estabelecido baseado em três etapas:

- **Etap 1 - Apresentação:** O principal objetivo desta etapa é a apresentação da estrutura e dos componentes do modelo proposto, junto aos requisitos necessários para o avanço de maturidade em Governança Ágil em TIC.
- **Etap 2 - Diagnóstico:** O principal objetivo desta etapa é executar um diagnóstico de maturidade, expondo forças, fraquezas e oportunidades de melhorias para a organização.
- **Etap 3 - Aperfeiçoamento:** O principal objetivo desta etapa é promover o alcance de níveis mais avançados de maturidade, baseado nos resultados da etapa anterior.

3.2 Níveis de Maturidade e Processos

Os processos especificados para o modelo foram construídos com base nas normas ISO/IEC 12207 e ISO/IEC 15504-2. Para tal, cada processo foi descrito em termos de um único propósito e de um conjunto de resultados esperados. As organizações geram os resultados esperados, conforme executam os processos, e estes resultados podem servir como indicadores de desempenho.

A estrutura básica definida para construção de cada processo do M3 foi estruturada da seguinte maneira:

- **Processo:** Conjunto de ações que se relacionam. Após execução destas ações é gerado um conjunto de resultados esperados predefinidos (produtos). Cada processo possui uma sigla e um nome que o identificam unicamente.
- **Propósito:** Define os objetivos do processo.
- **Resultados Esperados:** Produtos obtidos após a execução do processo. Produtos de trabalho ou mudanças significativas resultantes da execução de um processo os evidenciam.

O processo a seguir apresenta um exemplo prático de como esta estrutura foi utilizada para construção dos processos do M3:

Processo: Identificar e Priorizar Iniciativas (IPT)

Propósito: Analisar os objetivos estratégicos da área de TIC, identificando e priorizando iniciativas que promovam agilidade para um melhor alcance dos objetivos estratégicos organizacionais.

Resultados Esperados: As iniciativas de TIC que promovem agilidade são catalogadas em uma lista devidamente priorizada.

O modelo de maturidade proposto neste trabalho busca representar seu conhecimento em Governança Ágil em TIC através de níveis de maturidade. Estes níveis foram estruturados com base no conhecimento adquirido durante todo o processo de pesquisa. Como resultado, o M3 foi concebido mediante cinco níveis. Estes níveis variam de 1 (Inicial) ao nível 5 (Maduro). São estes:

- **Nível 1 - Inicial:** a organização não apresenta evidências claras de iniciativas relacionadas à Governança Ágil em TIC.
- **Nível 2 - Repetível:** a aplicação dos processos de apoio à Governança em TIC na organização começa a ganhar maior destaque no quesito agilidade. As responsabilidades para promover uma Governança Ágil em TIC são claras e os processos que a apoiam devem ser submetidos a uma monitoração e controle inicial.
- **Nível 3 - Intermediário:** os processos que apoiam a Governança Ágil em TIC tornam-se ainda mais prioritários organizacionalmente. O plano e a estratégia para promover uma Governança Ágil em TIC são seguidos e atualizados com grande regularidade em busca de uma organização mais adaptativa. A organização trabalha baseada em um plano, políticas e processos através de uma definição clara dos papéis e responsabilidades para início de uma Governança Ágil em TIC mais efetiva.
- **Nível 4 - Avançado:** os processos de apoio à Governança Ágil em TIC são consolidados. Neste nível, os processos de apoio a Governança Ágil em TIC passam a ser melhor padronizados e difundidos organizacionalmente. Além disto, estes processos devem ser claramente documentados e constantemente avaliados através de medições, sendo controlados estatisticamente. Estes processos também devem ser constantemente submetidos à melhoria contínua.
- **Nível 5 - Maduro:** Tem como objetivo o estabelecimento de melhoria contínua. Neste nível existe uma preocupação em prospectar, selecionar e avaliar tecnologias e novos paradigmas que promovam uma Governança Ágil em TIC ainda mais efetiva.

Tabela 2. Processos e Níveis do M3.

Níveis	Processos da Dimensão Governança em TIC	Processos da Dimensão Agilidade
1	Ausência de claras evidências de iniciativas para Governança Ágil em TIC.	
2	- Iniciar Diagnóstico do Ambiente e Capacitação da Equipe. - Identificar e Priorizar Iniciativas. - Preparar Cadeia Cliente Fornecedor. - Preparar os ANS. - Preparar e Homologar PIG.	- Gestão de Relacionamento com os stakeholders. - Gestão Ágil de Planejamento e Organização. - Desenvolvimento Ágil de Soluções.
3	- Analisar Ambiente Organizacional. - Alinhar Iniciativas com os Modelos de Apoio a Governança. - Executar Capacitação da Equipe. - Detalhar Serviços da CCF	- Realizar Gestão Adaptativa. - Gestão Ágil da Qualidade.
4	- Gerenciar Itens do Ciclo de Implantação.	- Monitoramento e Controle Ágil.
5	- Identificar e Planejar Melhorias.	- Identificar e Planejar Melhorias.

Quanto aos seus processos, o MR foi especificado a partir de influências dos meta princípios para Governança Ágil em TIC. Estes meta princípios foram identificados a partir de uma revisão sistemática de literatura (LUNA et al., 2014). Além disto, estes processos também receberam influências dos princípios e valores do *Manifesto for agile software development* (BECK et al., 2001), do COBIT 5 (ISACA, 2012) e da ISO/38500. A **Tabela 2** apresenta os processos e níveis construídos para o M3.

4. ANÁLISE DOS RESULTADOS

Para Caplan (1990), um *focus group* é a reunião de um grupo de pessoas com objetivo de avaliar conceitos e/ou problemas evidenciados. O objetivo principal de um *focus group* é identificar o sentimento de participantes sobre um determinado assunto, produto ou atividade. Os seus objetivos específicos variam de acordo com a abordagem de pesquisa adotada. Em uma pesquisa exploratória, o seu propósito é a geração de novas ideias ou hipóteses, estimulando o pensamento do pesquisador.

Este trabalho realizou uma sessão de *focus group* com o objetivo de avaliar a versão beta do modelo de maturidade proposto. Esta avaliação foi realizada envolvendo especialistas das áreas de Governança em TIC, agilidade e/ou maturidade. Estes especialistas foram convidados a participar do estudo através de solicitações formais via e-mail. Para o convite, foram previamente enviados: um resumo da proposta de trabalho e os pré-requisitos para participação na pesquisa. Entre estes pré-requisitos foram definidos que os participantes deveriam possuir:

- Pelo menos 2 (dois) anos de experiência em Governança em TIC ou agilidade;
- Pelo menos 2 (dois) anos de experiência em maturidade;
- Preferencialmente certificações nas áreas chave da pesquisa (agilidade, maturidade e/ou Governança em TIC);
- Preferencialmente experiência com desenvolvimento de pesquisas em áreas correlatas.

Tabela 3. Perfil dos participantes do estudo.

Participantes	Função/Certificações/Formação
PA1	Consultor em melhorias de processo - Implementador e Avaliador de modelos de qualidade - Auditor Líder ISO 9000 e ISO 29110. Implementador e Avaliador Líder inicial MPS.BR, apto a participar como equipe em avaliações CMMI. Com 12, 6 e 7 anos de experiência respectivamente em maturidade, Governança em TIC e agilidade. Doutorando com foco em gestão de projetos.
PA2	Consultor em melhorias de processo - Implementador e Avaliador de modelos de qualidade - Certificado ISTQB, implantador/avaliador MPT, implementador MPS.BR Serviços. Certificado Scrum Master (CSM), Certificado Scrum Developer (CSD). Criou um modelo de maturidade para Desenvolvimento Distribuído de Software em seu Doutorado. Com 5, 2 e 3 anos de experiência respectivamente em maturidade, Governança em TIC e agilidade.
PA3	Diretor de Governança e Tecnologia da Informação - Criou um modelo de maturidade para Planejamento Estratégico em seu Doutorado. Com 10, 6 e 5 anos de experiência respectivamente em maturidade, Governança em TIC e agilidade.
PA4	Gerente de Infraestrutura de TIC - Certificado Scrum Master (CSM), ITIL e COBIT. Desenvolve pesquisas em Governança Ágil desde 2008. Com 3, 12 e 13 anos de experiência respectivamente em maturidade, Governança em TIC e agilidade. Doutorando com foco em Governança Ágil.
PA5	Gerente de Projetos - Certificado ITIL e COBIT. Desenvolve pesquisas em Governança Ágil desde 2011. Com 2, 5 e 4 anos de experiência respectivamente em maturidade, Governança em TIC e agilidade. Mestre em gestão de projetos.
PA6	Gerente de Projetos - Certificado Scrum Master (CSM). Desenvolve pesquisas em gestão de projetos desde 2011. Com 2, 1 e 5 anos de experiência respectivamente em maturidade, Governança em TIC e agilidade. Doutorando com foco em gestão de projetos.

Definimos estes critérios buscando principalmente fazer com que tanto aspectos de Governança em TIC, como de agilidade e maturidade estavam sendo avaliados durante todo o estudo. Como resultado de nossos convites, seis (6) convidados puderam efetivamente colaborar com a pesquisa. Para tal, foi assinado um termo de confidencialidade para viabilização do estudo proposto. A **Tabela 3** apresenta o perfil dos participantes deste estudo.

Entre os participantes é possível identificar: um (1) Mestre, dois (2) Doutores e três (3) Doutorandos, todos em Ciência da Computação. É importante salientar que dois destes participantes têm atuado com uma média de aproximadamente seis (6) anos em Governança Ágil em TIC. Além disto, temos entre os participantes dois (2) consultores em melhorias de processo/implementadores e avaliadores de modelos de qualidade de referência no mercado, um (1) Gerente de Infraestrutura, um (1) Diretor de Governança em Tecnologia da Informação e dois (2) Gerentes de Projeto. Podemos ainda observar que os participantes do estudo possuem uma experiência média de aproximadamente seis (6) anos em maturidade, seis (6) em agilidade e cinco (5) em Governança em TIC. Estes especialistas foram submetidos a uma avaliação tanto da estrutura quanto dos processos do modelo proposto.

A definição sistemática de coleta de dados foi feita a partir de discussões moderadas pelo investigador proponente do modelo de maturidade para Governança Ágil em TIC. Para tal, inicialmente foi realizada uma breve apresentação do M3, visando tornar o método ainda mais eficiente. Em seguida, formulários foram distribuídos para cada um dos participantes contendo as informações que de fato deveriam ser avaliadas. Estes formulários foram estruturados para conter espaços reservados onde poderiam ser registradas observações relevantes para discussões durante o estudo. Além disto, estes formulários também foram utilizados posteriormente para uma análise mais detalhada das informações coletadas durante o estudo.

Antes mesmo do início das análises e preenchimentos destes formulários, foi então informado aos participantes do *focus group* que estes poderiam se sentir livres para expor opiniões diversas e agregando conhecimento, mesmo quando divergente das opiniões dos demais participantes. As seções a seguir apresentam de forma mais detalhada os resultados obtidos diante do estudo.

4.1 Análise das respostas: Estrutura do M3

O M3 foi concebido inicialmente sobre uma estrutura bidimensional, onde cada dimensão criada era composta por processos específicos. A primeira dimensão do modelo foi composta por processos que continham uma maior ênfase em Governança em TIC, enquanto que a segunda dimensão continha processos com uma maior ênfase em agilidade. A princípio, nenhum dos participantes do estudo esteve totalmente em desacordo com esta estrutura bidimensional, porém, após uma análise mais detalhada da estrutura do modelo proposto, foi então sugerida a fusão das dimensões em uma única estrutura. Sob a ótica dos participantes deste estudo, estas dimensões dificultavam tanto a compreensão quanto a operacionalização do modelo proposto.

Além disto, foi também sugerido que os processos do modelo fossem agrupados em categorias bem definidas, conforme outros modelos já o fazem (por exemplo, o COBIT 5 (ISACA, 2012) classifica seus processos em processos de Gestão e processos de Governança). Sob a ótica dos participantes do estudo, precede

aos processos uma estrutura que suporte um ecossistema que os abrigue e os organize. Isto melhora não só a estrutura do modelo proposto como também oferece uma maior clareza quanto ao foco de atuação de cada um de seus processos. Por fim, ainda no quesito estrutura, os participantes sugeriram que o modelo de maturidade proposto buscasse seguir o quantitativo de níveis dos modelos CMMI ou MPS.BR já que são modelos bastante consolidados atualmente no mercado.

Alguns dos comentários realizados pelos participantes do estudo foram:

- **PA1 comentou:** “é preciso avaliar o custo benefício de estruturar o modelo em duas dimensões. A estratégia bidimensional me parece aplicável para contextos mais simples, no entanto, a tentativa de tratar Governança em TIC e agilidade sob uma estrutura bidimensional pode gerar considerável complexidade”.
- **PA2 comentou:** “estruturar o modelo em dimensões me parece interessante, porém, no atual contexto aplicado, tais dimensões tem tornado o modelo um tanto quanto complexo desnecessariamente”.
- **PA3 comentou:** “a complexidade de uma estrutura bidimensional como a proposta na atual versão do modelo pode atrapalhar a compreensão e até dificultar a especificação e sinergia entre seus processos”.
- **PA4 comentou:** “me parece que operacionalizar de forma prática um modelo como este, baseado numa estratégia bidimensional, não é uma tarefa nada trivial. Isto foge do que é praticado por alguns dos modelos mais difundidos atualmente no mercado a exemplo do CMMI e do próprio MPS.BR”.
- **PA5 comentou:** “além da própria unificação desta estrutura bidimensional, é importante avaliar uma maneira de melhor organizar os processos. O COBIT 5, por exemplo, agrupa-os em processos de Gestão e processos de Governança. O CMMI e o MPS.BR também possuem seus próprios agrupamentos de processos”.
- **PA6 comentou:** “Uma boa estratégia para os níveis do modelo proposto seria seguir o quantitativo de níveis do CMMI ou MPS.BR já que são modelos bastante consolidados no mercado”.

4.2 Análise das respostas: Processos do M3

Com relação aos processos, nenhum dos especialistas discordou da importância geral destes para o contexto de Governança Ágil em TIC. No entanto, algumas melhorias envolvendo nomenclaturas, siglas e/ou propósitos, foram sugeridas com objetivo de aprimorar o modelo nos quesitos clareza e objetividade. Na visão dos participantes do estudo, as nomenclaturas, siglas e propósitos destes processos precisam ser autoexplicativos. Por exemplo, para o processo “Preparar Acordos de Nível de Serviço (PAS)” foi sugerido que este apresentasse uma nomenclatura/sigla como “Manter Acordos de Nível de Serviço (MANS)”. Além disto, seu propósito foi refinado para: “Manter os Acordos de Nível de Serviço (ANS) do Catálogo de Serviços de Tecnologia da Informação e Comunicação”. Outros refinamentos semelhantes envolvendo

nomenclaturas, siglas e/ou propósitos dos processos também foram discutidos e consolidados durante o estudo.

Durante uma análise mais detalhada dos processos, percebeu-se também que alguns destes estavam assumindo uma postura muito operacional e menos estratégica (principal característica da Governança). Uma reestruturação destes processos também foi alvo de discussões. Uma das questões levantadas neste sentido foi a inclusão de novos processos estratégicos para Governança em TIC (a exemplo de processos envolvendo atividades estratégicas relacionadas à riscos, transparência, financeiro, entre outras) e a eliminação de processos mais operacionais (a exemplo de processos como “Desenvolvimento Ágil de Soluções (DAS)”).

Por fim, além destes ajustes, também foram sugeridas algumas fusões entre os processos. Tais fusões tiveram como principal objetivo tornar o modelo mais simples e coeso, evitando assim redundâncias desnecessárias. Por exemplo, os processos “Gerenciar Itens do Ciclo de Implantação (GII)” e “Identificar e Planejar Melhorias (IPM)” tiveram sua unificação sugerida devido a sua semelhança em foco de atuação. Algo semelhante também foi discutido para os processos “Identificar e Priorizar Iniciativas de TIC (IPT)” e “Alinhar Iniciativas de TIC com Modelos de Apoio a Governança em TIC (AIG)”.

Alguns dos comentários realizados pelos participantes do estudo foram:

- **PA1 comentou:** “a princípio, todos os processos apresentam relevância, porém, alguns destes precisam ser refinados com objetivo de torná-los mais claros e objetivos em suas nomenclaturas, siglas e/ou propósitos”.
- **PA2 comentou:** “apesar da relevância dos processos para o contexto de Governança Ágil em TIC, alguns destes apresentam uma variedade excessiva de verbos em suas nomenclaturas e descrições de propósitos. São utilizados verbos como: Iniciar, “Analisar”, “Identificar”, “Executar”, entre outros. Seria interessante assumir um padrão”.
- **PA3 comentou:** “acredito que, mesmo com a relevância destes processos, alguns possam ser unificados, eliminados e outros incluídos, o que tornaria o modelo menos redundante, com maior foco no estratégico e com maior cobertura de atividades relacionadas à Governança em TIC”.
- **PA4 comentou:** “alguns processos tem assumido uma postura muito operacional. A meu ver, a Governança em TIC não deve fugir de sua característica estratégica”.
- **PA5 comentou:** “é importante existir uma relação direta entre as nomenclaturas/siglas de cada processo e seus respectivos propósitos. Neste sentido alguns processos precisam ser refinados visando evitar entendimento dúvida”.
- **PA6 comentou:** “Há processos que apresentam definições de elementos básicos para Governança em TIC na descrição de seus propósitos. Por exemplo, definições de elementos como: Acordo de Nível de Serviços (ANS), Portfólio de Serviços de Tecnologia da Informação e Comunicação (PSTIC), entre outros. Acredito que estas definições não devam fazer parte da especificação do propósito do processo”.

4.3 Análise das respostas: M3 em geral

Para os participantes do estudo, o M3 apresenta uma proposta interessante e desafiadora para o atual contexto de Governança em TIC. É preciso conciliar as exigências dos controles (internos e externos) organizacionais, com a necessidade por agilidade imposta por um mercado cada vez mais competitivo. O modelo proposto apresenta uma solução inovadora baseado em meta princípios que buscam conduzir a uma Governança mais ágil. Baseado em iniciativas de referência do mercado, o modelo proposto pode ser utilizado como uma alternativa de apoio a Governança Ágil em TIC. No entanto, diante das discussões apresentadas durante o *focus group*, podemos observar que o modelo requer aprimoramentos antes de sua aplicação prática em um ambiente real.

4.4 Trabalhos Relacionados e Contribuições

Alguns trabalhos relacionados foram identificados ao longo desta pesquisa. KORHONEN; YILDIZ; MYKKÄNEN (2009) apresentam uma das primeiras tentativas em criar um modelo de referência para a Governança de Segurança em SOA. LI; BIENNIER; GHEDIRA (2011) propõem um método de Governança Ágil para arquiteturas industriais multi-camadas, buscando atender aos princípios de Serviços de TIC (utilizando SOA e BPMN) combinados com as teorias *Lean* e *Six Sigma*. Para tal, foram definidos três níveis para o método: Negócios, serviços e aplicação.

Por outro lado, WILKINSON (2006) descreve uma estratégia para empresa adaptativa na HP. Neste trabalho são apresentados estágios de maturidade que tratam da típica evolução de uma empresa até seu estágio adaptativo. DOVE; HARTMAN; BENSON (1996) apresentam um modelo de maturidade para avaliar a competência de organizações em se adaptar diante de mudanças.

SUN; ZHANG; VALOTA (2005), por sua vez, propõem um conjunto de ferramentas que analisam e medem quantitativamente a agilidade de organizações de manufatura. YUSUF et al. (2004) utiliza um modelo conceitual que consiste em quatro dimensões para avaliar a maturidade de uma *agile supply chain*. Por fim, AMBLER (2009) define um modelo de maturidade para Governança de processos ágeis de desenvolvimento de software. Este mesmo autor afirma que a Governança não é algo que é comumente associada a projetos de desenvolvimento ágil de software, mas que, no entanto, os projetos ágeis, assim como todos os outros projetos, podem e devem ser governados.

Desta forma, baseado nas subáreas identificadas para Governança Ágil em nossa revisão sistemática de literatura (LUNA et al., 2014), podemos observar que os trabalhos apresentados abordam temas diversos: Os dois primeiros abordam SOA *Governance* (KORHONEN; YILDIZ; MYKKÄNEN, 2009; LI; BIENNIER; GHEDIRA, 2011); os dois seguintes *Enterprise Architecture* (DOVE; HARTMAN; BENSON, 1996; WILKINSON, 2006); e os restantes *Agile Manufacturing* (YUSUF et al., 2004) e *Governance for Agile Software Development* (AMBLER, 2009) respectivamente.

Apenas três destes trabalhos descrevem uma clara proposta para evolução de maturidade baseada em estágios, baseando-se em uma postura mais sistemática e gradual a mudanças. Neste sentido, é possível observar trabalhos atuando em: *Governance*

for *Enterprise Architecture*, *Agile Manufacturing* e *Agile Software Development*, mas não em *Agile IT Governance*.

Além disto, os trabalhos que envolvem TIC de alguma maneira, o fazem apenas sob um determinado aspecto: Segurança da Informação, Desenvolvimento de Software ou Serviços de TIC. Nenhum destes trabalhos apresenta uma proposta de Governança Ágil em TIC de maneira holística.

Levando em consideração estas lacunas, o M3 traz contribuições para a subárea de *Agile IT Governance*, apresentando um modelo que apoia a evolução sistemática e gradual de maturidade em Governança Ágil em TIC nas organizações, baseado em estágios de maturidade e seguindo uma abordagem holística.

5. CONCLUSÕES

Este artigo apresentou uma proposta de modelo de maturidade para Governança Ágil em TIC. Este modelo foi construído a partir de insumos importantes identificados durante todo o processo de pesquisa. O modelo construído é aderente às principais normas para construção de modelos de maturidade e Governança em TIC. Além disto, o modelo proposto foi estruturado a partir de um conjunto de meta princípios, identificados em uma ampla revisão bibliográfica da área. Por fim, este artigo também apresenta os resultados de uma avaliação do modelo proposto realizada por meio de *focus group*.

Os resultados desta avaliação têm contribuído para a evolução do modelo proposto, uma vez que tornou possível a obtenção de *feedbacks* de grande relevância a partir de profissionais ligados diretamente ao mercado. Durante este *focus group* foi possível identificar melhorias em diversos pontos do M3. Sugestões de melhoria referentes à sua estrutura, clareza/organização de seus processos (nomenclaturas, siglas e propósitos), quantitativo de níveis de maturidade, entre outras informações gerais envolvendo o modelo proposto, puderam ser discutidas e consolidadas.

Por fim, os participantes apontaram o modelo como promissor, ressaltando que este deveria ser colocado em prática em breve (em um ambiente real) para observação de seus reais benefícios no mercado. Entendemos que um modelo como o proposto neste trabalho de pesquisa, apresenta relevantes contribuições tanto para um contexto mercadológico quanto para o acadêmico.

Como trabalhos futuros, uma nova versão do modelo já está sendo concebida mediante os resultados obtidos nesta avaliação, a qual futuramente deverá ser submetida a um ambiente real para estudos de caso e avaliação de seus benefícios práticos.

REFERÊNCIAS

- ALMEIDA NETO, H. R. DE et al. **Uma Análise Comparativa envolvendo Modelos de Maturidade Ágeis**. Anais: XI - Simpósio Brasileiro de Qualidade de Software - VI Workshop de Desenvolvimento Rápido de Aplicações. 2012a.
- ALMEIDA NETO, H. R. DE et al. **A Comparative Analysis between Maturity Models for ITIL**. IV Congreso Internacional de Computación y Telecomunicaciones. Lima, Perú: 2012b. Disponível em: <<http://www.comtel.pe/comtel2012/callforpaper2012/P48C.pdf>>
- ALMEIDA NETO, H. R. DE et al. **A Comparative Analysis between Maturity Models based on COBIT**. III Encontro Anual de Tecnologia da Informação (EATI 2012). 2012c. Disponível em: <<http://eati.info/eati/2012/selecionados/>>
- ALMEIDA NETO, H. R. DE et al. **Evaluating a Maturity Model for Agile Governance in Information and Communication Technology**

- with Survey Based on Expert Opinion. 18 Workshop em Engenharia de Software Experimental (ESELAW 2015). Lima, Perú: 2015. Disponível em: <<http://eventos.spc.org.pe/cibse2015/sessions.html>>
- ALMEIDA NETO, H. R. DE; MOURA, H. P. DE. **MAnGve Maturity Model (M3): Proposing a Maturity Model to Support Agile Governance in Information and Communication Technology**. Anais: X Simpósio Brasileiro de Sistemas de Informação - Workshop de Teses e Dissertações (WTDSI). Londrina, Brasil. 2014. Disponível em: <<http://www.lbd.dcc.ufmg.br/colecoes/wtdsi/2014/009.pdf>>
- ALMEIDA NETO, H. R. DE; MOURA, H. P. DE; CAMPOS DE MAGALHAES, E. M. **MAnGve maturity model (M3): A proposal for a doctoral thesis**. Anais: 9th Iberian Conference on Information Systems and Technologies (CISTI). Barcelona, Espanha: 2014. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6876909>>
- AMBLER, S. **Scaling agile software development through lean governance**. Software Development Governance, 2009. SDG', p. 1–2, 2009.
- BECK, K. et al. **Manifesto for Agile Software Development**. Disponível em: <<http://agilemanifesto.org/>>. Acesso em: 25 abr. 2015.
- CADBURY, A. **The Financial Aspects of Corporate Governance**. The Committee on the Financial Aspects of Corporate Governance, UK, p. 90, 1992.
- CAPLAN, S. **Using focus group methodology for ergonomic design**. Ergonomics, v. 33, n. 5, p. 527–533, 1990.
- CARMEL, E. **The Offshoring Stage Model: an epilogue**, 2005. Disponível em: <auapps.american.edu/~carmel/papers/epilogue.pdf>
- DOVE, R.; HARTMAN, S.; BENSON, S. **An Agile Enterprise Reference Model with a Case Study of Remmele EngineeringForum or from Paradigm Shift International**, 1996. Disponível em: <<http://www.parshift.com/docs/aermodA0.htm>>
- GREMBERGEN, W. V.; DEHAES, S. D. **Implementing Information Technology Governance: Models, Practices and Cases**. New York, USA.: IGI Publishing, 2008.
- HEUSER, C. A. **Projeto de Banco de Dados**. 4. ed. Instituto de Informática da UFRGS., 1998. v. 4
- ISACA. **The Val IT Framework 2.0**, 2009. Disponível em: <http://www.isaca.org/knowledge-center/val-it-it-value-delivery-/pages/val-it1.aspx?utm_source=multiple&utm_medium=multiple&utm_content=friendly&utm_campaign=valid>
- ISACA. **COBIT 5: A Business Framework for the Governance and Management of Enterprise IT**. 2012.
- ITGI. **Board briefing on IT governance** 2nd ed. Rolling Meadows: IT. Governance Institute, , 2001.
- KORHONEN, J. J.; YILDIZ, M.; MYKKÄNEN, J. **Governance of Information Security Elements in Service-Oriented Enterprise Architecture**. Anais: 10th International Symposium on Pervasive Systems, Algorithms, and Networks, 2009. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5381732>>. Acesso em: 4 jan. 2012
- KTATA, O.; LÉVESQUE, G. **Agile development: issues and avenues requiring a substantial enhancement of the business perspective in large projects**. Anais: C3S2E '09 Proceedings of the 2nd Canadian Conference on Computer Science and Software Engineering. 2009. Disponível em: <<http://dl.acm.org/citation.cfm?id=1557636>>. Acesso em: 18 set. 2012
- LI, J.; BIENNIER, F.; GHEDIRA, C. **An Agile Governance Method for Multi-tier Industrial Architecture**. *liris.cnrs.fr*, p. 8, 2011.
- LOEWENSTERN, D. et al. **Dispatch tooling for global service delivery**. Anais: 2010 International Conference on Network and Service Management, 2010. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5691244>>
- LUNA, A. J. H. DE O. **MAnGve: Um Modelo para Governança Ágil em TIC**. Dissertação de Mestrado em Ciência da Computação. Programa de Pós-graduação stricto sensu em Ciência da Computação. Universidade Federal de Pernambuco. 2009.
- LUNA, A. J. H. DE O. et al. **Agile Governance in Information and Communication Technologies: Shifting Paradigms**. JISTEM Journal of Information Systems and Technology Management, v. 7, n. 2, p. 311–334, 2010.
- LUNA, A. J. H. DE O. et al. **State of the Art of Agile Governance: A Systematic Review**. International Journal of Computer Science and Information Technology, v. 6, n. 5, p. 121–141, 2014.
- LUNA, A.J.H. O., KRUCHTEN, P., MOURA, H. . **GAME: Governance for Agile management of enterprises a management model for agile governance**. Anais: Proceedings IEEE 8th International Conference on Global Software Engineering Workshops, ICGSEW 2013, p. 88–90, 2013.
- MARCONI, M.; LAKATOS, E. **Fundamentos de metodologia científica**. 2003.
- OGC. **ITIL v3 - Service Lifecycle - Introduction to ITIL**, 2009.
- PEDROSA, M. L. G. DO E. **MDIGA: Um Modelo de Perfis de indicadores para Governança Ágil em TIC**. Dissertação de Mestrado. Programa de Pós-graduação em Ciências da Computação. Universidade Federal de Pernambuco (UFPE), 2014.
- PRADO, D. S. **Maturidade em Gerenciamento de Projetos**. Nova Lima: INDG Tecnologia e Serviços Ltda, 2008.
- RATIONAL. **Rational Unified Process Best Practices for Software Development Teams** Rational, 2004. Disponível em: <<http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Rational+Unified+Process+Best+Practices+for+Software#4>>
- SEI. **Capability Maturity Model Integration, CMMI for Development (CMMI-DEV V1.3) CMU/SEI-2010-TR-033 ESC-TR-2010-033**. Pittsburgh: Disponível em: <<http://www.sei.cmu.edu/reports/10tr033.pdf>>.
- SOLER, A. **OPM3: a Contribuição PMI para Maturidade em Gestão de Projetos**. Revista Mundo PM, p. 46–49. v 1, n. 2. ISSN: 1807–8095, 2005.
- SUN, Y.; ZHANG, Z.; VALOTA, O. **A Methodology to form agile strategies in manufacturing organisations**. Anais: Proceedings IEEE International Engineering Management Conference, 2005, Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1559163>>
- WEIL, P.; ROSS, J. W. **IT Governance: How Top Performers Manage IT**. International Journal of Eletronic Government Research, v. 1, n. 4, p. 63–67, 2005.
- WILKINSON, M. (SPL-0045) [S115] **Designing an “adaptive” enterprise architecture**. BT Technology Journal, v. 24, n. 4, p. 81–92, 30 out. 2006.
- YUSUF, Y. Y. et al. **Agile supply chain capabilities: Determinants of competitive objectives**. European Journal of Operational Research, v. 159, n. 2 SPEC. ISS., p. 379–392, 2004.

Aplicação da otimização por colônia de formigas ao problema de múltiplos caixeiros viajantes no atendimento de ordens de serviço nas empresas de distribuição de energia elétrica

Alternative Title: Applying a variation of the ant colony optimization algorithm to solve the multiple traveling salesmen problem to route the teams of the electric power distribution companies

Denilson F. Barbosa
Programa de Pós-Graduação
em Informática
Universidade Tecnológica
Federal do Paraná
denilsonbarbosa@utfpr.edu.br

Carlos N. Silla Jr.
Programa de Pós-Graduação
em Informática
Universidade Tecnológica
Federal do Paraná
carlosjunior@utfpr.edu.br

André Y. Kashiwabara
Programa de Pós-Graduação
em Informática
Universidade Tecnológica
Federal do Paraná
kashiwabara@utfpr.edu.br

RESUMO

Este artigo apresenta uma metodologia para otimização do atendimento comercial nas empresas de distribuição de energia elétrica, atividade que representa uma parcela significativa dos custos operacionais dessas empresas. Como distribuir as ordens de serviço entre as equipes de atendimento e como definir rotas eficientes são problemas complexos devido à grande quantidade de combinações possíveis. Através da criação de instâncias de Problemas de Múltiplos Caixeiros Viajantes a partir das posições de execução das ordens e da aplicação de um algoritmo baseado na Otimização por Colônia de Formigas nessas instâncias, foram obtidas soluções otimizadas para distribuição das ordens e roteamento das equipes. O método desenvolvido foi aplicado a 17 instâncias construídas a partir de dados reais de uma agência de atendimento e mostrou-se eficiente ao diminuir em média 42,23% das distâncias percorridas pelas equipes em comparação com as rotas reais. Para possibilitar a replicação dos experimentos realizados, o código fonte do protótipo desenvolvido e os dados reais utilizados encontram-se disponíveis em <https://github.com/denilsonfag/STACS>.

Palavras-Chave

Distribuição de energia elétrica, problema de múltiplos caixeiros viajantes, otimização por colônia de formigas.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

ABSTRACT

This paper presents a methodology to optimize the execution of the commercial services of electric power distribution companies. This activity represents a significant portion of the operational costs of these companies. How to distribute the services to the teams and define efficient routes is a complex problem due to the great number of possible combinations. The set of geographical positions of the services can be seen as an instance of the multiple traveling salesmen problem. We created a variation of the ant colony optimization algorithm to obtain optimized solutions for the distribution of the services and the route of each team. Our methodology was applied to 17 real word instances obtained from an electric power distribution company from the city of Cornélio Procopio, Brazil. Our method obtained an improvement of 42,23% on average when compared to the solutions that were performed in the real world. To allow the reproduction of our results, the source-code of our system and the dataset are freely available at <https://github.com/denilsonfag/STACS>.

Categories and Subject Descriptors

I.2.8 [Artificial Intelligence]: Problem Solving, Control Methods, and Search—*heuristic methods*; G.1.6 [Numerical Analysis]: Optimization—*constrained optimization, integer programming, stochastic programming*

General Terms

Algorithms, Experimentation

Keywords

Distribution of electric power, multiple traveling salesmen problem, ant colony optimization.

1. INTRODUÇÃO

As principais atividades das empresas de distribuição de energia elétrica são a manutenção do sistema elétrico de distribuição e a gestão comercial de seus consumidores. Para a realização destas atividades, equipes técnicas percorrem diariamente a área sob concessão dessas empresas, o que representa uma parcela significativa de seus custos operacionais. Estes custos podem ser reduzidos com a aplicação de uma metodologia que utilize de forma eficiente as informações disponíveis para gerenciamento das equipes, como a apresentada neste artigo.

Em geral, a solução adotada nas empresas para gerenciamento do atendimento comercial é a execução dessa atividade por técnicos que realizam manualmente a análise das informações disponíveis ao distribuir as ordens entre as equipes no início dos dias de trabalho. Após distribuídas as ordens, cabe a cada equipe definir também manualmente a sequência em que realizará seus atendimentos. Tanto a distribuição das ordens quanto as sequências de atendimento admitem um grande número de combinações, o que as tornam tarefas difíceis de serem otimizadas sem o auxílio de uma ferramenta computacional.

A partir de informações obtidas da agência de atendimento de Cornélio Procopio, estado do Paraná, este trabalho apresenta uma metodologia para realizar a distribuição das ordens e criar rotas otimizadas para as equipes simultaneamente, através da aplicação de um algoritmo de otimização por colônia de formigas em instâncias de problemas de múltiplos caixeiros viajantes representando os dias de trabalho. Os experimentos realizados comprovam a eficácia da metodologia ao reduzir em 42,23%, em média, o deslocamento total das equipes em um dos experimentos realizados.

O restante deste trabalho está organizando da seguinte forma: na seção 2 são apresentados os trabalhos relacionados ao problema real e à abordagem deste artigo. Na seção 3 é detalhada a metodologia desenvolvida e na seção 4 são apresentados os experimentos realizados e os resultados obtidos. A seção 5 analisa a eficácia deste trabalho e aponta as questões a serem consideradas na sua continuação.

2. REFERENCIAL TEÓRICO

Na literatura, o trabalho de Costa et al. [3] aborda o problema da distribuição de ordens de serviço entre as equipes através da análise do histórico de atendimento de uma agência, buscando definir a melhor forma de particionar a área de atendimento para equilibrar a carga de trabalho das equipes. O problema é modelado como um Problema de p-Mediana Capacitado e é utilizada a Programação Linear Inteira e um Algoritmo Genético específico para tratá-lo.

A proposta nos trabalhos de Garcia et al. [7] e Mascia et al. [11] para distribuir as ordens entre as equipes também é através do Problema de Agrupamento Capacitado (outra denominação do Problema de p-Mediana Capacitado), só que os agrupamentos são realizados diretamente sobre as ordens a executar, definidas por suas posições. Nesta abordagem, após distribuídas as ordens entre as equipes, ainda cabe a cada uma delas definir manualmente a sequência em que realizará seus atendimentos. Uma maneira de distribuir as ordens entre as equipes e definir as sequências de atendimento simultaneamente é abordar o problema como um Problema de Múltiplos Caixeiros Viajantes, que é explicado na seção 2.1.

2.1 O Problema de Múltiplos Caixeiros Viajantes

O TSP (*Travelling Salesman Problem*, Problema do Caixeiro Viajante) é um dos problemas de otimização combinatoria mais pesquisados, que se destaca por ser de fácil declaração mas de difícil resolução, pertencendo à classe dos problemas NP-completos [1]. Em sua definição clássica, em uma instância do problema existem n cidades a serem visitadas por um caixeiro, que objetiva realizar a menor rota possível que passe por todas as cidades exatamente uma vez e retorne à cidade de partida. Várias aplicações do mundo real podem ser modeladas como instâncias de um TSP, como rotas de transporte escolar, sequenciamento de DNA, manufatura de circuitos de microchips e inspeções em plataformas de petróleo [9].

Uma instância de TSP pode ser representada em um grafo completamente conectado e não direcionado $G = (V, A)$, no qual V é o conjunto de n vértices representando as cidades e A é o conjunto de arestas que conectam todas as cidades entre si. Cada elemento de A possui um custo associado, que geralmente é uma medida de distância ou tempo, os quais são organizados numa matriz C na qual cada elemento c_{ij} é o custo do deslocamento iniciando no vértice $i \in V$ até o vértice $j \in V$. A matriz de custos é dita simétrica quando $c_{ij} = c_{ji}, \forall i, j \in V$, ou assimétrica caso contrário.

O MTSP (*Multiple Traveling Salesman Problem*, Problema de Múltiplos Caixeiros Viajantes) é uma generalização do TSP na qual mais de um caixeiro participa da solução. Como explica Bektas [2], o MTSP é mais apropriado para aplicações do mundo real que o TSP e pode ser estendido para uma grande variedade de VRPs (*Vehicle Routing Problem*, Problema de Roteamento de Veículos) com a adição de restrições às soluções. VRP é o nome dado a toda uma classe de problemas envolvendo a visita de clientes por veículos.

Na definição geral do MTSP, existem $m > 1$ caixeiros inicialmente posicionados em um vértice comum, normalmente indexado como o primeiro vértice da instância, $v_0 \in V$, e denominado depósito. Os demais vértices da instância são chamados de intermediários. O objetivo é minimizar o custo total da solução, que corresponde aos custos das rotas de todos os caixeiros somados, sendo que todas as rotas devem começar e terminar no depósito e todos os nós intermediários devem ser visitados exatamente uma vez. Neste artigo, o termo “rota” é empregado ao conjunto de cidades visitadas por um caixeiro ordenadas pelo instante da visita, e o termo “solução” é empregado para uma solução completa do problema.

Entre as formas possíveis de modelar o MTSP, a mais comum é a formulação baseada em atribuição utilizando a programação linear inteira, como a apresentada a seguir, extraída de [2]. Nessa formulação, n corresponde ao número de elementos de V , que é o conjunto de vértices da instância incluindo o depósito; m é o número de caixeiros; i e j são vértices da instância; x_{ij} é a variável binária que informa se a aresta (i, j) faz parte da solução do problema; c_{ij} é o custo associado à aresta (i, j) ; e S é o número de elementos dos subconjuntos de V , que constituem os vértices atribuídos a cada caixeiro. Na criação do modelo, inicialmente é definida a variável binária x_{ij} :

$$x_{ij} = \begin{cases} 1 & \text{se a aresta } (i, j) \text{ é usada na solução,} \\ 0 & \text{caso contrário.} \end{cases} \quad (1)$$

A função objetivo é:

$$\text{Min} \sum_{i=1}^n \sum_{j=1}^n c_{ij} x_{ij} \quad (2)$$

Sujeita às seguintes restrições:

$$\sum_{j=2}^n x_{1j} = m \quad (3)$$

$$\sum_{j=2}^n x_{j1} = m \quad (4)$$

$$\sum_{i=1}^n x_{ij} = 1, \quad j = 2, \dots, n \quad (5)$$

$$\sum_{j=1}^n x_{ij} = 1, \quad i = 2, \dots, n \quad (6)$$

$$\sum_{i \notin S} \sum_{j \in S} x_{ij} \geq 1, \quad \forall S \subseteq V \setminus \{1\}, \quad S \neq \emptyset \quad (7)$$

$$x_{ij} \in \{0,1\}, \forall i, j \in V \quad (8)$$

As restrições 3 e 4 garantem que exatamente m caixeiros farão parte da solução. As restrições 5, 6 e 8 representam as regras de atribuição e a restrição 7 previne a criação de subconjuntos de cardinalidade S que não contenham o depósito.

Uma variação importante do MTSP está relacionada ao objetivo do problema. Okonjo-Adigwe [12] aponta que o MTSP sem restrições não considera importantes parâmetros de problemas reais, entre eles o equilíbrio da carga de trabalho (*workload balance*) entre os caixeiros. Para buscar o equilíbrio entre os custos das rotas em um MTSP, ao invés de minimizar o custo total da solução, o custo a ser minimizado passa a ser o da maior rota da solução, que é composta por m rotas, uma de cada caixeiro.

2.2 A Otimização por Colônia de Formigas

A Otimização por Colônia de Formigas (*Ant Colony Optimization*, ACO) está entre as principais meta-heurísticas usadas para o TSP [9]. Ela tem alcançado resultados consistentes para problemas de otimização combinatória, recebendo esta denominação por ser baseada na maneira como colônias de formigas reais encontram rotas mais curtas até suas fontes de alimento. A partir do primeiro algoritmo ACO, denominado *Ant System* [5], que obteve resultados encorajadores para o TSP, a ACO tem sido aplicada com sucesso para vários problemas NP-difíceis de otimização combinatória [6].

Na ACO, as soluções são geradas num processo iterativo no qual agentes inteligentes, as formigas, percorrem todas as cidades de uma instância de TSP, por exemplo, escolhendo a cada iteração a próxima cidade de sua rota de maneira estocástica através da Regra de Transição de Estado (RTE). A RTE do *Ant System* é mostrada na equação 9:

$$p_{ij}^k(t) = \begin{cases} \frac{[\tau_{ij}(t)]^\alpha [\eta_{ij}]^\beta}{\sum [\tau_{ij}(t)]^\alpha [\eta_{ij}]^\beta} & \text{se } j \text{ é uma aresta permitida à } k \\ 0 & \text{caso contrário} \end{cases} \quad (9)$$

Na equação 9, que determina a probabilidade de um vértice ser escolhido por uma formiga, i representa o vértice no qual a formiga k se encontra, e j um dos possíveis vértices para o qual ela pode se deslocar. Os vértices permitidos à formiga são aqueles ainda não visitados até a fase atual da construção da solução. Cada formiga armazena os nós já visitados em sua memória, que é denominada como lista tabu. $\tau_{ij}(t)$ corresponde à quantidade de feromônio presente na aresta entre os vértices i e j no instante t . η_{ij} corresponde à visibilidade da aresta, que é definida pelo valor $1/c_{ij}$, onde c_{ij} é o custo de deslocamento da aresta. α e β são parâmetros que definem o peso da trilha de feromônio e da visibilidade, respectivamente, na escolha do próximo vértice pela formiga.

2.2.1 O Ant Colony System

O ACS (*Ant Colony System*, Sistema de Colônia de Formigas) é um dos vários algoritmos que surgiram a partir de aperfeiçoamentos no *Ant System* (AS), entre eles o *Elitist AS*, o *Ant-Q*, o *MAX-MIN AS* e o *Rank-based AS* [6]. O ACS difere do AS em três pontos principais: utiliza uma RTE mais agressiva; a RAF (Regra de Atualização de Feromônio) determina que a evaporação e o depósito de feromônio, ao final de cada ciclo, acontecem apenas nas arestas pertencentes à *best-so-far solution*, ou seja, a melhor solução encontrada até o momento em uma execução do algoritmo; e, a cada vez que uma formiga utiliza uma aresta numa solução, uma quantidade pré-definida de feromônio é removida da aresta, para aumentar a exploração de caminhos alternativos. Outra diferença importante é em relação ao número N de formigas que percorrem o grafo e constroem suas próprias soluções simultaneamente: enquanto no *Ant System* o melhor valor encontrado experimentalmente é $N = n$, para o ACS o melhor valor encontrado é $N = 10$.

$$j = \begin{cases} \text{argmax}_{l \in S_i^k} \{\tau_{il} [\eta_{il}]^\beta\}, & \text{se } q \leq q_0; \\ J, & \text{caso contrário.} \end{cases} \quad (10)$$

A RTE do ACS está mostrada na equação 10, na qual j representa a cidade escolhida por uma formiga k que se encontra no vértice i ao se movimentar. Esta equação é chamada de regra proporcional pseudo-aleatória, pois o parâmetro q_0 define a porcentagem das escolhas que serão feitas de forma determinística pelas formigas, onde $0 \leq q_0 \leq 1$. Sendo q uma variável aleatória uniformemente distribuída em $[0, 1]$ e atualizada a cada movimento, se $q_0 = 1$, todas as escolhas das formigas serão realizadas de forma determinística, pelo valor máximo de $\tau_{il} [\eta_{il}]^\beta$, onde $l \in S_i^k$ corresponde ao conjunto de cidades permitidas à formiga no movimento. Se, ao contrário, $q_0 = 0$, todas as escolhas serão realizadas de forma estocástica, pois J representa uma cidade escolhida através das probabilidades calculadas pela equação 9, com a única diferença que o parâmetro α foi excluído, ou seja, seu valor foi fixado em 1. O melhor valor definido experimentalmente é $q_0 = 0.9$, o que significa que 90% das escolhas serão determinísticas.

A RAF do ACS define que a atualização de feromônio ocorre em dois momentos: uma atualização global, ao final de cada ciclo do algoritmo; e uma atualização local, logo depois que uma formiga se move de uma cidade a outra, ou seja, inclui mais uma aresta na sua rota. A regra para Atualização Global de Feromônio (AGF) está apresentada nas equações 11 e 12, onde ρ é o coeficiente de persistência

do feromônio, cujo melhor valor encontrado para o ACS é $\rho = 0.1$. T^{bs} corresponde ao conjunto de arestas que fazem parte da melhor solução encontrada até o momento atual da execução do algoritmo e C^{bs} corresponde ao custo de T^{bs} .

$$\tau_{ij} = (1 - \rho)\tau_{ij} + \rho\Delta\tau_{ij}^{bs}, \quad \forall(i, j) \in T^{bs}, \quad (11)$$

$$\Delta\tau_{ij}^{bs} = 1/C^{bs} \quad (12)$$

A regra de Atualização Local de Feromônio (ALF) está apresentada na equação 13, na qual ξ é um parâmetro cujo melhor valor experimentalmente calculado é $\xi = \rho = 0.1$. τ_0 também é um parâmetro que, além de ser utilizado na atualização local de feromônio, também corresponde à quantidade de feromônio depositada em todas as arestas no início da execução do algoritmo. O melhor valor encontrado para τ_0 é calculado a partir do custo de uma solução criada aplicando um algoritmo do vizinho mais próximo à instância em análise. A fórmula para cálculo de τ_0 é mostrada na equação 14, na qual n é igual ao número de cidades da instância e C^{nn} é igual ao custo da solução obtido utilizando um algoritmo do vizinho mais próximo sobre a instância.

$$\tau_{ij} = (1 - \xi)\tau_{ij} + \xi\tau_0, \quad (13)$$

$$\tau_0 = 1/nC^{nn} \quad (14)$$

2.2.2 O Team Ant Colony Optimization

O TACO (*Team Ant Colony Optimization*, Otimização por Colônia com Equipes de Formigas) é o algoritmo desenvolvido por Vallivaara [13] que adapta o ACS para instâncias de MTSP. Isto é feito substituindo cada formiga do ACS, que gera soluções para o TSP, por um time de formigas formado por m elementos, onde cada formiga de um time corresponde a um caixeiro na construção de uma solução para MTSP, de forma que cada time possui sua própria lista tabu.

No início da construção da solução, todas as formigas de todos os times são posicionadas no depósito. Para realizar a distribuição da carga de trabalho, a formiga que tem a menor rota parcial, em qualquer momento da construção do solução, escolhe sua próxima cidade de acordo com a regra definida na equação 10. Segundo Vallivaara [13], este método é eficiente para a distribuição da carga de trabalho, mas em muitos casos as escolhas das formigas tendem para soluções não ótimas. Para tratar este problema, é verificado a cada escolha se, caso a cidade escolhida fosse cedida a qualquer outra formiga do time, não haveria um tamanho total menor da solução, considerando também o retorno ao depósito. Se houver, é a formiga com menor rota parcial que se movimenta, podendo realizar uma nova escolha do seu próximo vértice a partir da regra 10.

Matematicamente, estando uma formiga k posicionada em um vértice v_k ; sendo v_0 o depósito da instância; $c(v_i, v_j)$ o custo de deslocamento entre dois vértices v_i e v_j quaisquer; $CP(k)$ o custo parcial total da rota já percorrida pela formiga k ; e tendo sido escolhido o vértice v_j como seu próximo vértice de acordo com a regra 10; o método consiste em verificar, antes da formiga de mover, se existe qualquer outra formiga l da instância em que

$$c(v_l, v_j) + c(v_j, v_0) + CP(l) < c(v_k, v_j) + c(v_j, v_0) + CP(k) \quad (15)$$

Se houver, é permitido que a formiga l escolha seu próximo vértice novamente de acordo com a equação 10, e se movimente para o novo vértice escolhido.

Outra técnica utilizada por Vallivaara [13] é a aplicação de algoritmos de buscas locais para melhoramento das soluções geradas. A busca local 2-opt, que tenta trocar duas arestas quaisquer da solução, é aplicada a todas as soluções geradas. A busca local 3-opt, que tenta comutar quaisquer três arestas de uma solução ao mesmo tempo, é aplicada apenas à melhor solução dentre as N geradas simultaneamente pelos times. Para a otimização das instâncias de MSPT geradas a partir do problema real em estudo, foi desenvolvido um protótipo que implementa uma variação do TACO de Vallivaara [13], que é explicado na seção 3.

3. METODOLOGIA

A metodologia desenvolvida consiste na criação de instâncias de MTSP a partir de cenários estáticos do ambiente em estudo e na aplicação de um algoritmo ACO a estas instâncias para criação de soluções otimizadas para distribuição das ordens e roteamento das equipes. O ambiente analisado é o atendimento no município de Cornélio Procopio, estado do Paraná, que possui $637,95 \text{ km}^2$ de extensão territorial e 19.276 consumidores de energia elétrica, dos quais 3,76% estão localizados na área rural [8]. As equipes de atendimento ficam concentradas na mesma garagem e estão aptas a executar qualquer tipo de ordem de serviço, não havendo, portanto, restrições quanto à habilidade das equipes na distribuição das ordens.

Na figura 1 estão destacadas em vermelho as 31 ordens de serviço executadas no município em 3 de fevereiro de 2014, um dia de trabalho típico. O ponto em azul representa a garagem da agência de atendimento, local onde as equipes normalmente iniciam e concluem suas rotas. Nos eixos da figura 1 estão representadas as coordenadas no sistema de projeção Universal Transverso de Mercator (UTM), que definem as posições da garagem e das ordens executadas. O UTM é o sistema de coordenadas cartesianas bidimensional oficial brasileiro para mapeamentos [10]. Essas coordenadas foram obtidas do histórico da agência de atendimento, a partir do qual também foi possível obter a sequência em que as equipes executaram os serviços, o que tornou possível o cálculo das rotas reais realizadas.

Para criar e otimizar soluções para instâncias de MTSP como a da figura 1, foi desenvolvido na linguagem de programação C++ um protótipo que implementa um algoritmo ACO denominado STACS (*Single Team Ant Colony System*, Sistema de Colônia com Equipe Única de Formigas). O STACS é baseado no *Team Ant Colony Optimization* (TACO) [13], que por sua vez adapta o *Ant Colony System* (ACS) [4] para o MTSP. Esta implementação é apresentada no algoritmo 1. A principal diferença entre os dois algoritmos é que no TACO N soluções são geradas simultaneamente por N times, enquanto que no protótipo desenvolvido as soluções são geradas uma a uma e, quando N soluções são geradas, o que corresponde ao final de um ciclo, ocorre a atualização global de feromônio. Outra diferença é em relação à busca local 3-opt, que não é aplicada às soluções geradas pelo STACS devido ao tempo computacional exigido.

O protótipo desenvolvido aceita como entrada instâncias de MTSP representadas por coordenadas planas, como a da figura 1, para as quais são calculadas as respectivas matrizes de custos com as distâncias euclidianas entre os pontos;

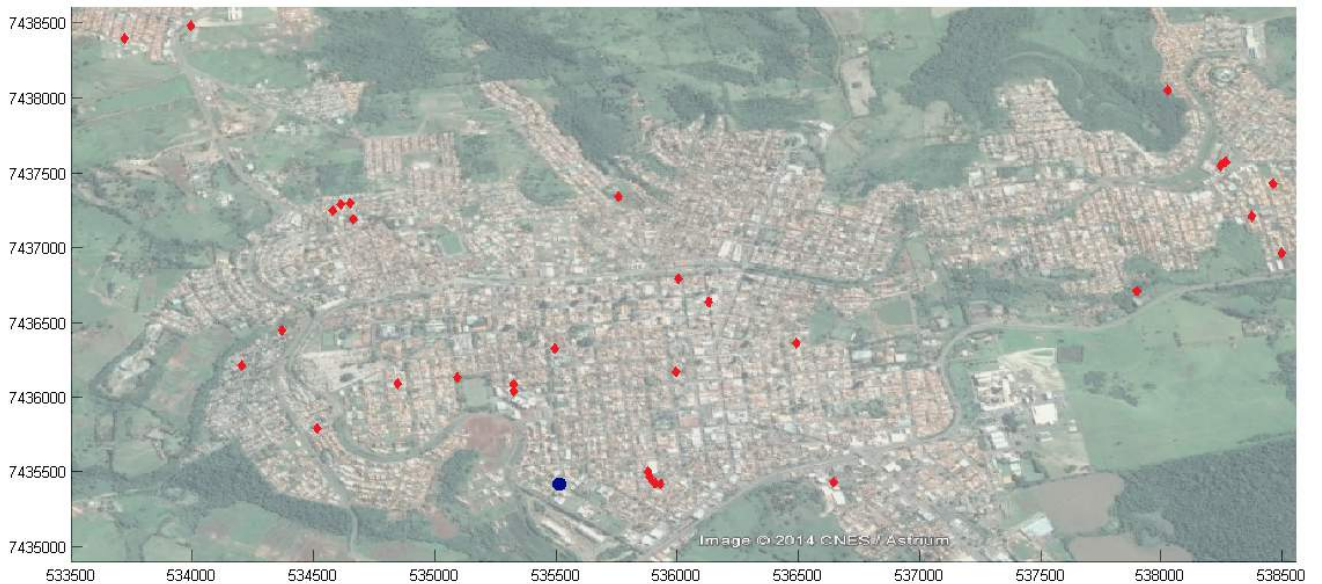


Figura 1: Ordens de serviço executadas no município de Cornélio Procopio em um dia de trabalho típico (pontos em vermelho) e a garagem da agência de atendimento (ponto em azul). Fontes: imagem de satélite de Google Maps e coordenadas dos pontos de COPEL Distribuição.

Algoritmo 1: STACS (*Single Team Ant Colony System*)

Entrada: Instância MTSP e parâmetros ACS

Saída: Melhor solução MTSP criada

```

1 Início
2    $m \leftarrow$  número de caixeiros da instância;
3    $n \leftarrow$  número de cidades da instância;
4    $v_0 \leftarrow$  índice do vértice depósito da instância;
5    $N \leftarrow$  número de soluções geradas em cada ciclo;
6   Inicialize a matriz de feromônio; //equação 14
7   Enquanto o critério de parada não for atingido faça
8     Para contador_de_soluções  $\leftarrow 1$  até  $N$  faça
9       Esvazie a lista tabu da colônia;
10      Posicione  $m$  formigas em  $v_0$ ;
11      Insira  $v_0$  na lista tabu;
12      Enquanto houver vértices não visitados faça
13         $k \leftarrow$  formiga com menor rota parcial;
14         $v_j \leftarrow$  próximo vértice de  $k$ ; //equação 10
15        Verifique o movimento de  $k$ ; //inequação 15
16        se houver uma formiga  $l$  melhor que  $k$  então
17           $k \leftarrow l$ ;
18           $v_j \leftarrow$  próximo vértice de  $k$ ; //equação 10
19        fim
20         $v_i \leftarrow$  vértice atual de  $k$ ;
21        Movimente a formiga  $k$  para o vértice  $v_j$ ;
22        Insira  $v_j$  na lista tabu;
23        Aplique ALF à aresta  $(v_i, v_j)$ ; //equação 13
24      Fim
25      Movimente todas as  $m$  formigas para  $v_0$ ;
26      Monte a solução MTSP criada em  $S_{atual}$ ;
27      Aplique a busca local 2-opt à  $S_{atual}$ ;
28      Atualize a melhor solução do ciclo  $S_{m\_ciclo}$ ;
29    Fim
30    Atualize a melhor solução da execução  $S_{m\_exec}$ ;
31    Aplique a AGF à  $S_{m\_exec}$ ; //equações 11 e 12
32  Fim
33  Retorne  $S_{m\_exec}$ ;
34 Fim

```

e instâncias representadas por matrizes de custo assimétricas, o que permite a utilização de custos que representem melhor o ambiente real da instância em análise, como tempos estimados de deslocamento, por exemplo. A distância euclidiana também é chamada de distância métrica e, para dois pontos i e j representados por coordenadas planas x e y , seu valor é igual a $[(x_i - x_j)^2 + (y_i - y_j)^2]^{1/2}$ [5]. Nos experimentos apresentados na seção 4 são utilizadas matrizes com as distâncias euclidianas nas otimizações.

4. EXPERIMENTOS COMPUTACIONAIS

Os experimentos apresentados nesta seção compravam a eficácia da metodologia desenvolvida aplicada ao problema real. Utilizando dados de uma agência de atendimento, são realizados dois experimentos. No primeiro, o protótipo é configurado para minimizar o custo total das soluções, sem considerar a distribuição da carga de trabalho entre as equipes, como na descrição geral do MTSP. No segundo experimento, o protótipo minimiza o custo da maior rota individual das soluções, objetivando a construção de soluções formadas por rotas com custos equilibrados entre os caixeiros, como no MTSP com equilíbrio da carga de trabalho. Para possibilitar a replicação destes experimentos, em <https://github.com/denilsonfag/STACS> encontram-se disponíveis o código fonte do protótipo desenvolvido e os dados reais utilizados.

4.1 Detalhes experimentais

As ordens de serviço da base de dados reais foram separadas quanto ao seu dia de execução e, a partir das coordenadas UTM que indicam suas posições de execução, foi montada uma instância de MTSP para cada dia de trabalho. Além das coordenadas das ordens, as instâncias contêm as coordenadas da agência de atendimento (obtidas através de uma consulta no Google Maps a partir de seu endereço)

e o número de equipes em serviço no dia, que corresponde ao número m de caixeiros.

Para a confecção das matrizes de custo que representam as instâncias, foram utilizadas as distâncias euclidianas, as quais foram calculadas a partir das coordenadas UTM dos pontos divididas por 1000, o que gera valores em quilômetros, já que cada unidade no sistema de projeção UTM corresponde a um metro na superfície da Terra [10].

4.1.1 Base de dados reais

A base de dados reais utilizada nos experimentos consiste no histórico de atendimento do dia 1º de fevereiro a 23 de fevereiro de 2014 do município de Cornélio Procopio. Os dados foram obtidos a partir do sistema de gerenciamento das ordens de serviço da COPEL Distribuição que, dentre outras funções, armazena as informações sobre as ordens desde a sua geração até a sua conclusão. Os atendimentos realizados pelas equipes nos 23 dias está resumido na tabela 1, que traz a divisão das ordens de acordo com seu dia de execução e o número de equipes em serviço em cada dia. Nos dias 1º, 2, 8, 9 e 23, havia apenas uma equipe em serviço, o que gera instâncias para TSP, pois $m = 1$. Como o foco deste artigo é o MTSP, estes dias de trabalho não foram utilizados nos experimentos, apesar de isto ser possível no protótipo desenvolvido. No dia 16 não houve nenhum atendimento, sendo também excluído da base de dados. Dessa forma, os experimentos foram realizados sobre os 17 dias de trabalho restantes.

Tabela 1: Ordens de serviço executadas no município de Cornélio Procopio de 1º de fevereiro a 23 de fevereiro de 2014. Fonte: COPEL Distribuição.

Dia de trabalho	Equipes	Ordens de serviço
1	1	7
2	1	1
3	3	31
4	4	38
5	4	40
6	4	33
7	4	40
8	1	4
9	1	1
10	3	37
11	4	27
12	4	53
13	4	35
14	4	33
15	2	12
16	0	0
17	2	26
18	4	64
19	4	35
20	4	30
21	4	36
22	2	9
23	1	5

A partir do histórico de atendimento foram obtidas as coordenadas UTM de todas as ordens executadas, que possibilitam a construção de instâncias de MTSP representando os 17 dias de trabalho. Analisando o horário de execução das ordens, também disponível no histórico, foi possível definir a sequência real dos atendimentos nos dias de trabalho, que corresponde a uma solução para MTSP, cujos custos calculados pelo protótipo com a utilização das distâncias eucli-

dianas servem de referência para a avaliação dos resultados obtidos e estão mostrados nas tabelas 2 e 4.

4.1.2 Protocolo experimental

Em cada um dos experimentos foram realizadas 100 execuções independentes do algoritmo STACS para cada uma das 17 instâncias. O algoritmo foi implementado em C++ e compilado com o pacote de desenvolvimento *Cygwin 1.7.32*. Foi utilizado um equipamento de 64 bits executando o sistema operacional *Windows 8.1* em um processador com dois núcleos de 2,40 GHz e 8 GB de memória RAM. As sementes dos números pseudo-aleatórios de todas as execuções foram armazenadas no arquivo de *log* do experimento para possibilitar a replicação dos resultados.

O critério de parada do algoritmo STACS foi de 1000 iterações em cada execução. O conjunto de parâmetros utilizado foi : $q_0 = 0.9$, $\alpha = 1$, $\beta = 2$ e $\xi = \rho = 0.1$. Foram geradas $N = 10$ soluções a cada ciclo. Esses valores foram os melhores obtidos experimentalmente por [6] ao aplicar o ACS a um grande conjunto de instâncias euclidianas de TSP e também foram os mesmos utilizados por [13] em seus experimentos.

4.2 Minimização do custo total das soluções

O primeiro experimento executado com os dados e o protocolo descritos foi realizado minimizando-se o custo total das soluções, ou seja, a soma de todos os custos das rotas individuais das equipes. Isto significa que, nas comparações entre duas soluções durante a execução do algoritmo, a que é considerada como sendo melhor é a que tiver menor custo total. Esta configuração, que pode ser útil no ambiente real caso se deseje exclusivamente reduzir o deslocamento total das equipes, sem se preocupar com a distribuição equilibrada das ordens entre elas. Os resultados obtidos estão sintetizados na tabela 2.

4.2.1 Tempos de execução do algoritmo

A tabela 3 apresenta o tempo médio de execução do algoritmo para cada instância. Este valor foi obtido a partir da média do tempo gasto para execução dos 1000 ciclos do STACS nas 100 execuções do algoritmo, para cada dia de trabalho.

Tabela 3: Tempo médio de execução de 1.000 ciclos do STACS para cada dia de trabalho.

Dia de trabalho	Número de equipes	Número de ordens de serviço	Tempo médio de execução (ms)
3	3	31	2794
4	4	38	3805
5	4	40	4426
6	4	33	3327
7	4	40	4368
10	3	37	3534
11	4	27	2414
12	4	53	7197
13	4	35	4010
14	4	33	3667
15	2	12	677
17	2	26	1846
18	4	64	9993
19	4	35	3448
20	4	30	2658
21	4	36	3629
22	2	9	658

Tabela 2: Resultados obtidos minimizando-se o custo total das soluções MTSP para os 17 dias de trabalho.

Dia	Solução Real		Médias de 100 execuções do STACS				Melhoramento	
	Custo Total (CTr)	Maior Rota (MRr)	Custo Total (CTp)	D. P.	Maior Rota (MRp)	D. P.	Custo Total (CTr - CTp)	Maior Rota (MRr - MRp)
3	60,83	22,37	19,53	0,29	13,86	2,67	41,30	8,51
4	84,41	41,39	52,03	0,38	39,45	5,01	32,38	1,94
5	131,98	84,36	94,00	0,44	75,17	4,56	37,98	9,19
6	116,16	54,24	75,24	1,14	52,71	8,05	40,92	1,53
7	109,53	51,15	68,94	0,46	43,83	6,80	40,59	7,32
10	62,22	23,47	21,36	0,27	15,36	3,38	40,86	8,11
11	48,95	18,61	18,87	0,35	15,44	2,58	30,08	3,17
12	188,19	74,20	131,72	0,74	115,40	9,41	56,47	-41,20
13	52,25	23,05	21,01	0,32	13,43	1,51	31,24	9,62
14	60,53	23,78	23,07	0,44	14,42	3,73	37,46	9,36
15	140,44	133,77	111,08	0,01	104,84	0,00	29,36	28,93
17	37,75	20,22	16,75	0,46	13,26	3,44	21,00	6,96
18	71,73	24,05	27,13	0,26	15,32	3,89	44,60	8,73
19	56,26	21,65	26,09	0,60	19,61	1,91	30,17	2,04
20	55,31	23,83	22,23	0,43	12,05	3,40	33,08	11,78
21	70,62	28,70	24,22	0,43	15,55	3,32	46,40	13,15
22	138,59	108,06	105,03	0,00	103,05	0,00	33,56	5,01
Totais	1485,75	776,90	858,29		682,75		627,46	94,15

4.3 Minimização da maior rota individual das soluções

Um segundo experimento foi realizado minimizando-se a maior rota individual das soluções e mantendo-se os demais parâmetros como no primeiro experimento. Esta configuração é adequada ao ambiente real caso se deseje que as equipes tenham deslocamentos equilibrados entre si ao final da execução das ordens, o que também resulta que mais ordens sejam executados no mesmo intervalo de tempo (uma equipe não fica parada devido a ter uma rota significativamente menor que as outras, por exemplo). A tabela 4 apresenta os resultados obtidos para esta configuração.

4.4 Análise dos Resultados

O STACS apresentou resultados satisfatórios para as duas variações de MTSP experimentadas: minimizando o custo total da solução (tabela 2) e minimizando o custo da maior rota individual da solução (tabela 4). Os pequenos valores dos desvios padrão nas duas tabelas confirmam a robustez do algoritmo ao gerar soluções com custos próximos à média das 100 execuções.

No primeiro experimento, houve uma diminuição de 627,46 km no deslocamento total das equipes nos 17 dias de trabalho, comparando os custos reais com a média dos custos das soluções obtidas nas 100 execuções do STACS, o que corresponde a uma melhora de 42,23% em relação às rotas reais das equipes. Além disso, para todas as 17 instâncias foi encontrada uma solução com custo total menor e apenas para o dia de trabalho 12 houve um aumento da maior rota individual da solução.

No segundo experimento, configurado para minimizar a menor rota individual, a redução do deslocamento total das equipes nos 17 dias foi de 294,82 km, significativamente inferior ao ganho obtido no experimento anterior. Porém, quando verificando a maior rota individual das equipes, tem-se uma melhora de 94,15 km no primeiro experimento e 269,90 km no segundo. Apesar do ganho total nos deslocamentos ser menor, o ganho nas rotas individuais informa que as equipes trabalharam de maneira mais uniforme, con-

cluindo suas rotas em instantes próximos, o que leva à execução de mais ordens em menos tempo, aumentando a qualidade do atendimento prestado.

Quanto aos tempos de execução do algoritmo, mostrados na tabela 3, verifica-se que para a maior instância experimentada, a do dia 18, com 64 ordens de serviço, foram gastos em média aproximadamente 10 segundos para execução dos 1000 ciclos do algoritmo, que representam um tempo aceitável para o problema em estudo.

5. CONCLUSÕES

A metodologia apresentada mostrou-se eficiente na distribuição das ordens de serviço entre as equipes e na criação de rotas otimizadas para realização dos atendimentos. Entretanto, algumas questões precisam ser analisadas, como a utilização de distâncias euclidianas para representação dos custos de deslocamento reais das equipes. A continuação deste trabalho prevê a utilização de custos que representem de forma mais fiel o ambiente real, como previsões de deslocamento entre dois pontos fornecidas por Sistemas de Informações Geográficas.

Quanto à confirmação da eficácia do método, é necessária a comparação dos resultados apresentados neste artigo com outros obtidos a partir da aplicação de diferentes metodologias ao problema. Para a criação de soluções otimizadas para MTSP neste trabalho, foi implementado um algoritmo baseado na meta-heurística ACO. Outra meta-heurística que tem obtido resultados consistentes para problemas de otimização combinatória são os Algoritmos Genéticos (AG) [3], que serão utilizados para comparação com os resultados obtidos pelo STACS em novos experimentos.

Quanto ao conjunto de parâmetros do algoritmo ACO utilizado nos experimentos, que foram obtidos a partir de trabalhos relacionados, uma experimentação utilizando o STACS e instâncias construídas a partir de dados reais podem resultar em um conjunto de parâmetros mais eficiente para o problema em estudo. Porém, devido à grande quantidade de combinações possíveis de parâmetros, esta experimentação assume elevada complexidade e será realizada na continua-

Tabela 4: Resultados obtidos minimizando-se o custo da maior rota individual das soluções MTSP para os 17 dias de trabalho.

Dia	Solução Real		Médias de 100 execuções do STACS				Melhoramento	
	Custo Total (CTr)	Maior Rota (MRr)	Custo Total (CTp)	D. P.	Maior Rota (MRp)	D. P.	Custo Total (CTr - CTp)	Maior Rota (MRr - MRp)
3	60,83	22,37	24,16	0,54	8,32	0,03	36,67	14,05
4	84,41	41,39	61,80	2,38	33,89	0,00	22,61	7,50
5	131,98	84,36	103,94	1,60	73,29	0,00	28,04	11,07
6	116,16	54,24	127,12	12,25	39,58	0,08	-10,96	14,66
7	109,53	51,15	80,79	1,91	39,81	0,00	28,74	11,34
10	62,22	23,47	24,96	0,26	8,39	0,11	37,26	15,08
11	48,95	18,61	28,15	1,54	7,73	0,00	20,80	10,88
12	188,19	74,20	185,48	18,73	60,80	0,00	2,71	13,40
13	52,25	23,05	32,16	2,26	9,26	0,02	20,09	13,79
14	60,53	23,78	30,55	0,47	7,90	0,02	29,98	15,88
15	140,44	133,77	179,35	1,54	90,21	0,67	-38,91	43,56
17	37,75	20,22	17,49	0,10	8,77	0,06	20,26	11,45
18	71,73	24,05	32,00	0,41	8,14	0,09	39,73	15,91
19	56,26	21,65	38,77	1,58	11,64	0,02	17,49	10,01
20	55,31	23,83	28,90	0,21	7,41	0,02	26,41	16,42
21	70,62	28,70	32,44	0,14	8,15	0,02	38,18	20,55
22	138,59	108,06	162,86	3,74	83,70	0,00	-24,27	24,36
Totais	1485,75	776,90	1190,93		507,00		294,82	269,90

ção deste trabalho.

Outro ponto diz respeito ao caráter dinâmico do ambiente em estudo. Enquanto a metodologia proposta mostra-se viável para cenários estáticos, no qual todas as ordens e as posições das equipes são conhecidas antes da otimização, no ambiente real serviços surgem durante o dia de trabalho das equipes e podem ter maior prioridade que outros, como as ordens emergenciais. Para tratamento desta questão, a metodologia desenvolvida será adaptada para gerar novas soluções otimizadas a cada mudança no ambiente real, como o surgimento de novas ordens.

6. AGRADECIMENTOS

À Alexandre da Silva, técnico de atendimento comercial e emergencial da COPEL Distribuição na agência de Cornélio Procopio, pelo fornecimento dos dados reais utilizados nos experimentos.

7. REFERÊNCIAS

- [1] J. Augustine. *Offline and online variants of the traveling salesman problem*. PhD thesis, Louisiana State University, Department of Electrical and Computer Engineering, 2002.
- [2] T. Bektas. The multiple traveling salesman problem: an overview of formulations and solution procedures. *Omega*, 34(3):209–219, June 2006.
- [3] C. E. d. S. Costa, D. M. B. Costa, and A. R. T. Goes. Determinação de setores de atendimento em uma concessionária de energia. In *Anais do Simpósio Brasileiro de Pesquisa Operacional - SBPO*, pages 951–962, Goiânia, 2006.
- [4] M. Dorigo and L. Gambardella. Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, 1(1):53–66, Apr. 1997.
- [5] M. Dorigo, V. Maniezzo, and a. Colorni. The Ant System: optimization by a colony of cooperating agents. *IEEE transactions on systems, man, and cybernetics.*, 26(1):29–41, Jan. 1996.
- [6] M. Dorigo and T. Stutzle. *Ant Colony Optimization*. The MIT Press, Cambridge, Massachusetts, 2004.
- [7] V. Garcia, D. Bernardon, M. Sperandio, C. do Vale, and J. Fernandes. Service order dispatching in electric utilities. In *45th International Universities Power Engineering Conference (UPEC)*, pages 1 – 7, Cardiff, Wales, 2010. IEEE.
- [8] IPARDES. Caderno estatístico do município de Cornélio Procopio. *Instituto Paranaense de Desenvolvimento Econômico e Social*, 2014.
- [9] J. Y. Kanda. Sistema de meta-aprendizado para a seleção de meta-heurísticas para o problema do caixeiro viajante. In *X SBSI - Simpósio Brasileiro de Sistemas de Informação*, pages 651–662, Londrina, 2014.
- [10] D. M. Machado Jr and M. A. Dal Santo. Considerações acerca de Trabalhos em Áreas de Divisa de Fusos UTM. In *Anais do Congresso Brasileiro de Cadastro Técnico Multifinalitário*, pages 1–14, 2004.
- [11] A. R. Mascia, W. d. M. Reck, V. J. Garcia, D. P. Bernardon, M. Sperandio, and C. do Vale. Algoritmo heurístico para agrupamento de ordens de serviço em concessionárias de distribuição de energia elétrica considerando priorização. In *Anais do Congreso Internacional de Distribucion Electrica*, pages 1–5, Buenos Aires, 2010.
- [12] C. Okonjo-Adigwe. An effective method of balancing the workload amongst salesmen. *Omega*, 16(2):159–163, Jan. 1988.
- [13] I. Vallivaara. A team ant colony optimization algorithm for the multiple travelling salesmen problem with minmax objective. In *Proceedings of the 27th IASTED International Conference on Modelling, Identification and Control*, pages 387–392, Anaheim, CA, USA, 2008.

Learning Difficulties in Computing Courses: Cognitive Processes Assessment Methods Research and Application

Vinicius Vieira Pessoni¹, Fernando Marques Federson², Auri Marcelo Rizzo Vincenzi³

Instituto de Informática
Universidade Federal de Goiás - UFG
Goiânia - Goiás, Brazil

¹viniciuspessoni@ufg.br, ²fernandofederson@inf.ufg.br, ³auri@inf.ufg.br

ABSTRACT

Learning difficulties in computing courses is a situation perceived in diverse universities from different countries, cultures and backgrounds. These difficulties directly affect achievement rates and increase course evasion. We believe in the existence of a foundation of cognitive processes, that without it, even the most motivated student would have trouble to transform the received information into knowledge. This work has focused mainly on the research of candidate methods for cognitive processes assessment with a strong background theory. With this kind of information would be possible to devise cognitive interventions, in order to evolve students cognitive level, and consequently, raise their success rates. A systematic review was conducted and among the many researched methods we selected Lawson Classroom Test of Scientific Reasoning – LCTSR. Authorized by its author, we conducted the first translation of LCTSR to Brazilian Portuguese and administered to students of three undergraduate computing courses: Information Systems, Computer Science and Software Engineering. We also present results of its administration that we consider important to reinforce the above suggested strategy.

Categories and Subject Descriptors

K.3.2 [Computers and Education]: Computer and Information Science Education – Information systems education

General Terms

Teaching, Learning, Cognition, Measurement, Human Factors.

Keywords

Computing Learning Difficulties, Cognitive Development, Formal Operational Stage, Scientific Reasoning.

1. INTRODUCTION

Student's learning difficulties in computing courses is not a recent, isolated and local preoccupation. Instead, it is a challenge perceived across the years in diverse universities from different countries, cultures and backgrounds [15, 28, 45, 52, 54]. These difficulties can be noted on student's low grades, achievement rates and course retention. They also influence negatively on course's engagement [6].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.
Copyright SBC 2015.

Despite it is relatively new, the Computer Education Research (CER) is going beyond to only propose tools to support learning computing and is gaining attention and importance as a wide research field [46, 51]. We agree with Malmi et al. [45] that the Computer Science, where theories are few and the constructive design and formal approaches dominate, has not favored the extensive use of existing theory as a normal part of research, like in the natural and human sciences. Therefore, we share the vision that a more holistic understanding of CER, beyond our basic approach, is necessary not only to consolidate its identity and independence as a research area, but also to achieve long term results.

There are many available tools to improve teaching and learning experiences. Through the years, the authors have been studying and applying many of them like: Bloom's taxonomy [31], as a classification of the different learning objectives but also as a guide to students about the cognitive process of learning in association with Kolb's learning cycle [30]; Myers-Briggs Type Indicator (MBTI), to measure psychological preferences in how people perceive the world and make decisions [9]; Kolb's Learning Style [27] and Index of Learning Styles (ILS) [17, 18], to measure the way individuals process information or prefer to learn. Some methodologies have also been used, like: Problem-based Learning (PBL) [3, 21] and Peer Instruction [14, 48], among others. There have also been some collaborative efforts from other groups in order to mitigate these issues, for instance [15, 47].

All these tools and methodologies are known by our scientific community. However, the deep understanding of their background theories has been of extremely importance for the authors to move their researches toward the comprehension of how knowledge construction is established on human beings, especially on those who left infancy phase. This way, even though the knowledge construction partially depends on factors like how the information is transmitted, student motivation, personality and interest, we believe that there is an inner foundation of cognitive processes, that without it, even the most motivated student would have difficulties to transform information into knowledge.

The analysis of anatomical and physical foundations of learning and memory is one of the greatest achievements of modern neuroscience. Thirty years ago little was known about "what are the different types of memory?", "where in the brain is memory located?" or "how memory is retrieved to answer a question?". Surely nowadays, knowledge is advancing and new findings are published every day [11, 49].

Although the researchers did not present a single or uniform theoretical viewpoint, it seems that the current research on learning and memory could be brought under a unifying

theoretical framework. This framework can be used to build and test metacognition models and strategies [66]. In this scenario, there is less room for intuitive models. Simons and Chabris [61], for example, shown that substantial numbers of their survey respondents agreed with propositions that conflicts with expert consensus about memory. They also emphasized that incorrect beliefs about the memory properties have broad implications, including the students misunderstanding of the memory role in learning processes.

Despite our approach is not precisely new, we believe it is relevant, being necessary for the community to revisit the previous existent models supported by neuroscience refreshing point of view. Especially in our country, Brazil, where teaching and learning indicators are not satisfactory, this kind of research would be a valuable contribution in order to develop strategies to improve teaching and learning processes.

The first step in this direction would be the identification of some kind of cognitive processes assessment method with a strong background theory. Then, with this kind of instrument, its model and some experience, it would possible to devise and apply approaches to verify the possibility of acquisition and development of cognitive processes on young and adults (metacognition).

The present work has focused mainly on the research of candidates methods to support this mission. We expect to be possible not only identify, characterize and measure cognitive processes (levels), but also to offer approaches to acquire and develop them. Therefore, reducing students gap among their cognitive level and the one expected by the courses, increasing the students success on these courses.

Among many researched methods, we choose and translated Lawson Classroom Test of Scientific Reasoning [34]. We also show results of its administration considered important to reinforce the above suggested strategy.

2. METHODOLOGY

The research was conducted through the following phases: 2.1. Cognitive Processes Assessment Methods Review; 2.2. Translation and Pilot Testing of LCTSR; 2.3. Application of the Translated Method. These phases are detailed in the next subsections. Results and data analysis are detailed on section 3.

2.1 Cognitive Processes Assessment Methods Review

A systematic review [29] was conducted in order to identify and select appropriate methods to assess University alumni cognitive processes. We researched relevant works in thirteen well settled computing and interdisciplinary digital libraries: ACM Digital library; IEEEExplore; ISI Web of Knowledge; ScienceDirect; Scopus; APA Psychnet; EBSCO (Academic Search Premier; CINAHL; SocINDEX; Library, Information Science & Technology Abstracts; Information Science & Technology Abstracts (ISTA); Academic Search Elite; Computers & Applied Sciences Complete; MEDLINE Complete).

In accordance with the objectives, we aimed to find instruments with a solid theoretical foundation, that would allow not only to identify, characterize and measure cognitive processes but would give us a more accurate information in order to develop approaches to generate permanent improvements on them. Among the instrument's background theories, Piaget's cognitive development [24] is one of the most solid and popular, being inspiration for many assessment methods.

Since Piaget developed his stage theory in the 50's [24, 53], its influences in teaching and learning science has pervaded numerous scientific productions, in multiple countries and in proportions that cannot even be measured. Therefore, today, many decades later, his insights are still very relevant in many ways [32].

Piaget [24], describes the cognitive development assessment method called "the clinical method" that consists of interviewing a person while presenting her a set of experiments with specific apparatus. Based on the person's answers it should be possible to classify her on the following cognitive development stages: sensorimotor, preoperational, concrete operations and formal operations [26]. Formal operations would be the stage in which the person would be more capable to deal with complex modes of reasoning, or in other words, abstraction.

Despite the researchers' preference for this method, the difficulties on applying it in large groups are listed as motivation for existence of almost all the posterior works that intends to perform Piaget's developmental level identification e.g [1, 34]. Some of recurrent critics on the clinical interview are the necessary expertise of the interviewer, the excessive space and time consumed, and the variation on influence of human factors in results, or lack of standardization.

The need of instruments that would be standardized, reliable and relatively simple to be administered to a large number of participants culminated in diverse evolutions and adaptations of the clinic interview, each one with its particularities. These adapted instruments differ in many aspects like: place of origin, presentation method, application field and distinction among cognitive stages.

The researches of methods are not restricted to one culture, country or continent, and each of them has its own values and contributions. Among these contributions, some notorious examples across the world are Shayer [60] in United Kingdom, Longeot [41] in France, Lawson [34] in United States of America, Bond [7] in Australia and Roadrangka [56] in Thailand.

The variations in presentation method occur in a way that some of them involve specific apparatus and experiments manipulation, and some do not. The experiments may be executed either by students themselves on their own kits or presented by the test administrator in front of class [4, 20, 44, 58, 60]. The apparatus and experiments required are basically the ones used by Piaget in the clinical method or some variations that are claimed to measure the same principles.

Still about the presentation variations, methods that do not use special apparatus or experiments and are fully described in paper, are generically known as pure or strict paper and pencils methods. These instruments are presented on questionnaire format, where the subject must read some breath paragraphs describing the question and sometimes analyze some graphical representation like pictures or diagrams. The answer sometimes open ended, other times objective, must be developed by himself, without test administrator intervention in the discovering process [1, 7, 10, 13, 41, 55–57, 59, 63–65].

In application field, the assessment methods may vary according to the age range of individuals. Some tests are adapted and validated for children [55], in others, adolescents age range are included [13, 19, 34, 63]. In a third variation are those that may be suitable from children to adults [4, 44, 56, 57, 64].

If the method is able to identify in which developmental stage individuals belong, it may be also a factor of distinction among tests. On the one hand, there are tests intended only to assess the presence of formal thinking, identifying some reasoning

patterns of this stage [4, 55, 64]. On the other hand there are tests that make a better distinction between formal operational and concrete, or previous stages, categorizing the subject in one of the four developmental stages [34, 56, 63].

Despite the previous variations, some others are described in [4]. For the present research, we preferred tests that could be administered to adolescents and adults, once its application is expected to be for University students, minimal age of 15 years old. Despite Shayer's [60] method have a great influence on scientific productions, it is preferable tests that do not involve specific apparatus manipulation and that could be objective scored. It is also desirable to distinct between formal operations and concrete or previous stages, to obtain some more complete information.

Thus, summarizing the criteria used for evaluating the assessment methods we expect that: 1. The method has a solid background theory; 2. The method does not involve any special apparatus manipulation; 3. Turns possible to clearly distinct the formal operational stage from previous stages; 4. Is suitable for adolescents and adults, people older than 15 years old; 5. Is objective scored, with no open ended questions.

According to these five criteria, after the systematic review, three methods of cognitive development assessment were selected as strong candidates for administration on our University: Arlin Test of Formal Reasoning – ATFR [1], Bond's Logical Operations Test – BLOT [7] and Lawson Classroom Test of Scientific Reasoning – LCTSR [34, 35]. Despite ATFR and BLOT demonstrated to be strong candidates, LCTSR has many advantages when compared to them. Some of these advantages are the massive cross-cultural previous applications and mainly the inclusion of the evolutions of cognitive theory through the years. This way, this exam was selected to be the first translated and used to assess alumni cognitive development on our University. The LCTSR is commented in the following section.

2.1.1 Lawson Classroom Test of Scientific Reasoning – LCTSR

The LCTSR is a paper and pencil instrument released in 1978. On its original version was comprised of 15 questions and some kind of demonstrations. In this format, the test administrator performs a demonstration in front of the class and poses a question. Students note their answers in test booklets, without sharing answers.

The instrument went through successive revisions and improvements across the years, with its most recent version released August 2000 [35]. In its current version, the exam is a pure paper and pencil instrument, with 24 two-tier items. Each of the test items require for the subject to select, from multiple alternatives and multiple justifications, in other words, the correct answer and justification for it. Scoring may be single or two tier type. This version has been designed to assess five reasoning modes of formal reasoning: controlling variables, proportional reasoning, combinatorial reasoning, probabilistic reasoning, and correlational reasoning. In addition, it has been extended to measure the hypothetical-deductive reasoning pattern, identifying the post-formal stage, of neo-piagetian theories [36].

Based on the student score on LCTSR, he is classified in one of the following cognitive levels: concrete (0-8); transitional (9-14); formal operational (15-20); post-formal (21-24).

Although some critics exist [8, 20, 42], LCTSR is wide validated and very popular among Science, Technology, Engineering, and Mathematics (STEM) educators and researchers not only in USA [16, 23]. It has been massively administered in several universities and countries, to students from different

graduation courses e.g. [2, 25, 43] including computing majors [50]. Due its strong validity and consistency, it has even become inspiration for many other tests [12, 33, 56, 64]. Its use is related not only to the intention to predict academic success, which is been concluded to be a valid predictor, but to develop materials and classes for cognitive interventions, helping students to improve their cognitive reasoning skills. We are particularly interested in this kind of use.

2.2 Translation and Pilot Testing of LCTSR

Since none of the researched methods were available in Brazilian Portuguese, translation and adjustments were necessary. After selecting LCTSR and verifying that no Portuguese version of it was available, the present authors communicated electronically with its author, Professor Lawson, requiring for permissions to translate and apply his test in our University. Once obtained the permissions, best practices on translation and methods for conducting it were researched.

To conduct a translation, especially of some assessment method, is necessary to follow a protocol to ensure that the translated method will be equivalent and assesses the same characteristics that the original does. Simple one direction translation, made by a single translator, is not recommended because the risks of translator bias and other personal inconsistencies that can be inserted during translating process [62]. This way, translation and cross-cultural adaptation must be carried out in a systematic process, following well defined phases.

In order to keep the translated instrument as equivalent as possible to the original, in the present research, the translation was conducted according to international well settled translation guidelines [5, 62] comprising the five following steps:

1) Independent Forwarding Translation: the English source test (ST) was translated by two bilingual independent translators (T1) and (T2). T1 was an informed translator, which means that he was familiar with the instrument background theories and objectives. T2 was an uninformed translator, meaning that he was not familiar with the method background theories neither its objectives. The results of this step were two translated distinct versions in Brazilian Portuguese (V1) and (V2).

2) Third Person Review: versions V1 and V2 were revised by a third bilingual person (T3), graduated in American Literature. V1 and V2 were compared with each other, and with the ST. The inconsistencies were noted in each translation, for further discussion in the next step. The results of this step were two revised and annotated distinct versions of the questionnaire (RV1) and (RV2).

3) Committee Review and Merge: the revised versions RV1 and RV2 were revised in a committee approaching, comprising four persons: the first independent translators T1 and T2, the T3 specialist reviewer, plus one of the present authors. Translation's divergences were discussed and resolved in group consensus, generating the merged candidate version (C).

4) Pilot Experimentation: the candidate version C, was then tested in a small group of voluntaries to verify translation and application inconsistencies. This group was very heterogeneous, with 25 subjects, with age range from 22 to 57, undergraduate, graduate and non undergraduate. Although there was no time limitation to accomplish the test, it was asked to voluntaries to annotate their start and end time, in order to measure how long they would take to answer it. The information of how much time they spent was used to calibrate the amount of time given to subjects on subsequent applications. The test was anonymous.

They were asked about the terminology used on the test and, if they had any difficulties on understanding about what was requested in each question. The participants' doubts on terminology and in any other aspect of the test were noted for further revision.

5) Final Revision and Version: the pilot group feedback was revised and some few mistakes on C were resolved. After including these adjustments, the final version (F) was generated. This final version was the one administered to computing courses alumni.

2.3 Subjects and LCTSR Application

After concluding the translation process, we administered the final version of the questionnaire to 210 subjects from seven different groups. 187 of these subjects were computing courses' students and 23 were from a special group. The computing courses' students comprise the following six groups:

- a) Information Systems First Semester (ISF);
- b) Information Systems Last Semester (ISL)
- c) Computer Science First Semester (CSF);
- d) Computer Science Last Semester (CSL);
- e) Software Engineering First Semester (SEF);
- f) Software Engineering Last Semester (SEL);

The seventh group was named "Outstanding Group" (OG). It is a very heterogeneous group formed by non-undergraduate, undergraduate and graduate people. The OG participants are different from the computing courses participants. Included in this group are people that popularly have been excelled, in some way, when compared to their colleagues. For example, they have the highest grades among their classmates, hold a high disputed position job or are known by its colleagues as the "smartest", just to name a few, from others perceptions of how they stands out.

The present authors administered the tests during regular classes. The test administrator presented himself in front of class and rapidly explained the objectives of our research. After that, he distributed the questionnaires to students who had fifty minutes to answer it. The fifty minutes' time was the calibrate amount of time concluded in the end of the pilot experimentation.

During application, if any student had doubts about what was being asked in a question, the test administrator answered him in front of class, to all the classmates to hear, according to the test recommendations. The students responded the test without discussion or sharing answers with classmates. All students finished to answer before the deadline.

Simple scoring was used, following Professor Lawson suggestions. To each correct answer, one point was assigned, ranging from 0 to a maximum of 24 points. According to this scoring system, the higher the score, the higher would be the students' abilities on scientific reasoning (formal reasoning).

3. Results and Data Analysis

The administration of a questionnaire like LCTSR to seven different groups allows the use of diverse analytical techniques. For example, the results can be obtained by analyzing all answers of some group (as a set), or only observing the differences between groups or even considering individual answers.

We highlight in the following subsections a few results we consider important for the objectives of this work. We also emphasize the authors' preference for nonparametric statistical methods because of, among other reasons, their reliance on fewer assumptions, what make them more robust and relatively insensitive to outlying observations [22].

3.1 The Outstanding Group

The Outstanding Group is known by its colleagues as the "smartest", as mentioned before. The need of such group in our work had two main objectives. The first objective was to verify if for the members of this group would be easier to achieve the highest score on the LCTSR. If it had happened, despite the questionnaire has its validity already verified by Lawson and other works, we would have to discard its use.

The results show that although 47.8% of this group are identified as post-formal reasoners and 52.2% as formal operational, only 2 of 23 participants has achieved a perfect score of 24 points. In addition, all participants reported that despite the questions were clearly stated and easy to understand, they were not easy to be solved. Therefore, we consider the instrument and its translation valid, in particular, with respect to the ability to distribute groups of our population in function of their correct answers.

The second objective of the OG was to be a reference to identify if the test is able to differentiate groups of students. In other words, this group is used to verify if the test is sensible enough to distinguish groups regarding their cognitive levels. The OG scores provide an upper limit or a challenging goal to be achieved by cognitive interventions.

3.2 First Semester versus Last Semester

The comparison between groups of first and last semester students from an undergraduate course gives us the possibility to verify whether, on average (as a group), it is possible to note a change in students cognitive ability during the course. The basic hypothesis is that, even as a group, students "should" develop (acquire, improve) their cognitive skills throughout the course.

Comparing all the answers of the groups SEF and SEL using the Wilcoxon-Mann-Whitney Test (WMW) [22], we obtained a p-value = 0.5748 ($W=679$), far above the traditional value of p-value < 0.05, so we can consider that the two groups belong to the same population. This way, when comparing these two groups, we cannot prove a change of cognitive abilities throughout the Software Engineering course. The percentage of individuals in each cognitive level proposed by Lawson, in the respective order (SEF, SEL) was: (22%, 20%) post-formal, (56%, 57%) formal operational and (22%, 23%) transitional.

The same conclusions can be drawn from the Computer Science course results. Comparing all the answers of the groups CSF and CSL, we obtained a p-value = 0.6442 ($W = 679$) and a distribution in the following order (CSF, CSL): (20%, 15%) post-formal, (48%, 55%) formal operational and (33%, 30%) transitional.

For the Information Systems course, the ISF group showed a lower percentage of individuals in the post-formal level (4%), but still we could achieve a p-value higher than 0.05, p-value = 0.0856 ($W = 91$) when compared to the ISL results. These results may be partially explained by the timing that the test was administered to the ISF group, very after another teaching activity of about one hour long. Differently, the test administration to the other groups were performed as the first academic day activity. Therefore it is important to be concerned with the way the test is administered (voluntary, in a relaxed and comfortable environment), the timing (preferably as the first academic activity) and duration (limited, but sufficient to avoid stress). Currently we have scheduled the retest of groups, spending more attention to these observations. The students' distribution following the order (ISF, ISL) was: (4%, 18%) post-formal, (62%, 64%) formal operational and (35%, 18%) transitional.

Despite the common sense, the results of LCTSR application do not permit us to affirm that the first semester and last semester groups of any courses show significant differences in their cognitive abilities.

3.3 Computing Courses

Here, we are interested in conduct comparisons among the three courses: Information Systems, Computer Science and Software Engineering. When comparing all the student's groups of the first semester, we tried to understand whether students who chose different courses would have distinct cognitive abilities. When comparing the student's groups of the last semester, we tried to identify in which cognitive level, on average, each course delivered to society their undergraduate students.

We used the Kruskal-Wallis test (KW) [22]. Thus we can decide whether the population distributions are identical without assuming them to follow the standard distribution. If p-value is less than 0.05, we can affirm that our collection of data samples are independent, so they come from unrelated populations.

Comparing the results of the questionnaire among the first semester students (SEF, CSF and ISF), we obtained a p-value = 0.0957 (KW chi-squared = 4.6937, df = 2). Figure 1 shows the box-plot for the three groups. Clearly the ISF group is responsible for a lower p-value. But still, we cannot claim to be found significant differences among populations regarding their cognitive abilities.

For the last semester students of the three courses (SEL, CSL and ISL) we obtained a p-value = 0.5360 (KW chi-squared = 1.2473, df = 2). Therefore, we cannot affirm that our last semester students, as a group, are concluding their graduation with differences in cognitive abilities in function of their courses. Figure 2 shows the box-plot for the three groups.

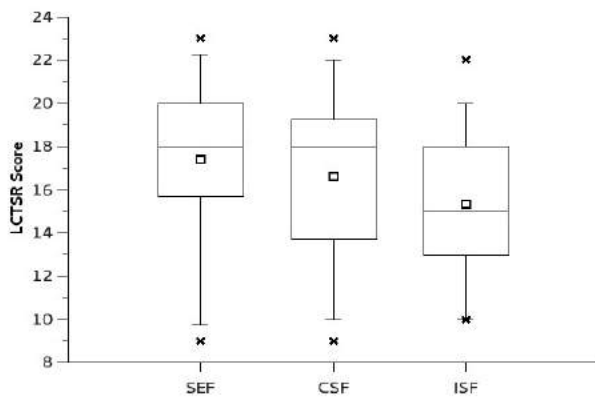


Fig. 1. Box-Plot: First Semester Students Comparison.

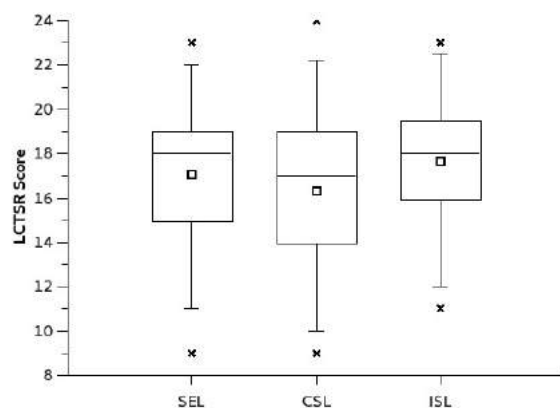


Fig. 2. Box-Plot: Last Semester Students Comparison.

3.4 The Outstanding Group versus Last Semester Groups

Until here, we cannot affirm that first and last semester students of the same course belongs to unrelated populations. Moreover, our results did not reveal cognitive differences among students of different courses. With these results, it is reasonable to question, for example, if the approach of treating students as a group is not very drastic, in the sense that it will be difficult to perform any intervention (conducted in a way that includes and influences all students from a group). The authors are aware of this great challenge, but Lawson [37] presents encouraging results where the interventions were statistically significant in a short period.

The comparison of the OG and the last semester students (SEL, CSL and ISL) shows that as a group, the OG belongs to a different population, with p-value = 0.0007 (KW chi-squared = 16.9068, df = 3). In an interview with members of the OG, it is evident that the vast majority of them assume to have a life story of particular taste for puzzle games and they also feel challenged by new knowledge. The authors believe that perhaps one of our greatest challenges is exactly to be able to provoke moments of new knowledge transmission in conjunction to the construction of cognitive skills. To carry out these interventions, the authors needed an instrument to measure the effectiveness of the intervention and LCTSR seems to be a strong candidate to this purpose.

3.5 Reasoning Patterns

The LCTSR consists of a sequence of questions, where each question or group of questions measure some kind of cognitive ability. Analyzing the mistakes of each researched group, a pattern was identified. Figure 3 shows the *relative frequency of wrong answers (rf)* given by two groups (SEF and SEL) on each question. The *rf* consists of the ratio between the total amount of errors of some question by the number of participants in the questionnaire (in percentage). Based on the *rf* profile it is possible to verify in which reasoning abilities the researched group has greater deficiencies and to devise approaches for an intervention.

The results reveal a pattern between the errors in the pair of questions 11-12 and 13-14. Surprisingly, the same pattern is also recognized in the results of all the other researched groups, including the OG (the pattern is highlighted in Fig. 4).

In order to confirm the unexpected relationship of the *rf* profiles among the researched groups, we computed the Spearman Rank Correlation Coefficient (r_s) [22] between all groups of results. Table 1 shows the values of r_s for each pair. All the table values are higher than the reference value of $r_s = 0.642$, for an alpha < 0.0005 (or 99.9995% of confidence) and n = 24. In other words, our results revealed a common error sequence. Based on Lawson cognitive level classification, it was expected that the *rf* would increase as the order of the question increases. This way, it was expected that the *rf* of the last items of the questionnaire would be greater than the *rf* of the first items.

After discovering this error pattern, double checking the translation quality and error counting, we conclude that this is really a significant result that can inspire and guide other researchers. Once LCTSR is proved to be able to identify the reasoning patterns in which students have more trouble, and their relation, the intervention possibilities are numerous.

Analyzing the cognitive skills that the pairs of questions 11-12 and 13-14 measure, according to Lawson [38], it is possible to verify that both pairs measure the "identification and control of variables and probabilistic thinking". Despite each of these skills

were separately assessed by the pairs of questions 9-10 and 15-16, respectively, apparently in these questions they were not so challenging for the most groups (this is confirmed by their lower rf). The results suggest that the need to use these two skills at the same time to solve question pairs 11-12 and 13-14, has been somehow a great challenge for our students (confirmed by the higher rf).

Once obtained these results, the authors started to develop educational materials in order to conduct a cognitive intervention. With this intervention we expect to help students to improve their cognitive levels. Acting specifically in the reasoning skills that LCTSR detected the higher rf , we hope to reduce the difficulties presented, improving the students success rates on courses.

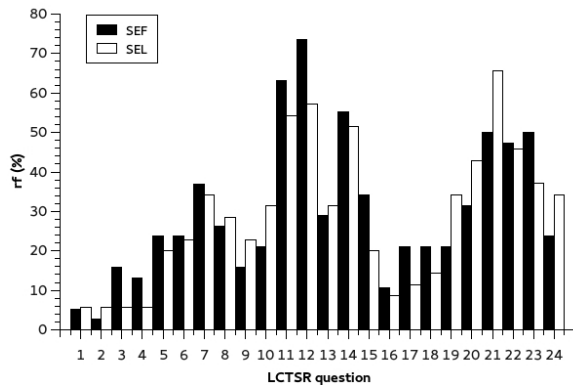


Fig. 3. Relative Frequency of Wrong Answers (SEF and SEL groups).

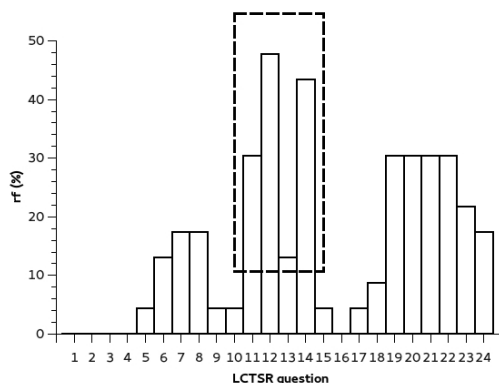


Fig. 4. Relative Frequency of Wrong Answers (OG group).

Table 1. Spearman Rank Correlation Coefficients

	CSF	CSL	SEF	SEL	ISF	ISL
OG	0.797	0.876	0.848	0.933	0.786	0.775
CSF	-	0.882	0.815	0.768	0.698	0.686
CSL	-	-	0.891	0.862	0.866	0.794
SEF	-	-	-	0.873	0.925	0.891
SEL	-	-	-	-	0.819	0.854
ISF	-	-	-	-	-	0.857

4. CONCLUSIONS

It is not difficult to find colleagues who, like the authors, believe that when students acquire new knowledge related to their expertise area, their cognitive levels change, in other words, increase with it. Lawson [39, 40] helps us with the differences between declarative knowledge and procedural knowledge:

“The word 'learning' is often used in conjunction with the acquisition of declarative knowledge. ...scientific reasoning consists of an overall pattern of reasoning, which can be characterized as hypothetic-deductive, as well as several sub-patterns the very process of generating and testing ideas results not only in the construction of ideas that work (i.e., the learning of useful declarative knowledge), but also in improved skill in learning (i.e., the development of improved procedural knowledge). Thus, to teach most effectively, teachers should allow their students to participate in the idea generation and testing.”

The present authors believe that the first step in order to develop strategies to improve teaching and learning processes would be the identification of instruments for cognitive processes assessment with a strong background theory.

In this work, it was found numerous methods, created through the years, in diverse countries and in many languages. As each one claims to be the best alternative, there was a need to organize them in a systematic way, and apply some criteria to help to decide which one to use. Of the many reviewed methods, ATFR, BLOT e LCTSR demonstrated to be strong candidates. None of them had Portuguese versions, this way, LCTSR was selected to be the first translated and administered to computing courses alumni (Information Systems, Computer Science and Software Engineering).

LCTSR was also administered to a special group in order to test the scale quality of the instrument and to verify if it would be sensible enough to distinguish students groups. The questionnaire and its translation proved to be valid for the initial objectives of the research, offering a good option to the scientific community.

The results were extremely encouraging, providing much more than just the classification of students in levels of cognitive ability. They allowed some initial work hypotheses to be tested, like the variation of cognitive abilities throughout the courses and the existence of differences among first and last semester students between courses. Based on our results, we cannot affirm that exist differences between first and last semester students of some course, neither of first and last semester among courses. The results also indicated which reasoning skills need more attention when conducting an intervention in order to be more efficient in the teaching process.

The present research brings uniqueness in two ways: 1. the first Brazilian Portuguese version of LCTSR is the one translated by the present authors; 2. the first application data and results of this test in our country.

Our future steps are: expand the application of the test, monitor over time the groups already tested and propose, according to the background theory, possible interventions.

5. ACKNOWLEDGMENTS

We would like to thank Professor Anton E. Lawson for providing the permission to translate and use his method. Professor Trevor G. Bond for his sympathy and professional cordiality on having shared part of his work. We would also like to thank Mr Fabrício Vieira Personi for his translation services and Miss Nayara Cristina Rodrigues de Andrade for her flawless revision.

6. REFERENCES

- [1] Arlin, P. K. 1982. A multitrait-multimethod validity study of a test of formal Reasoning. *Educational and Psychological Measurement*. 42 (1982), 1077 – 1088.
- [2] Ates, S. Cataloglu, E. 2007. The effects of students' reasoning abilities on conceptual understandings and

- problem-solving skills in introductory mechanics. *Eur. J. Phys.* 28 (2007), 1161–1171.
- [3] Barrows, H. S. Problem-based learning in medicine and beyond: A brief overview. *New directions for teaching and learning*. 68 (1996), 3-12.
- [4] Bart, W. M. 1972. Construction and Validation of Formal Reasoning Instruments. *Psychological Reports*, 30 (1972), 663 – 670.
- [5] Beaton, D. E. et al. Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures. *SPINE* 25, 24. (2000), 3186 –3191.
- [6] Bergin, S. Reilly, R. 2005. Programming: factors that influence success. In *Proceedings of the 36th SIGCSE technical symposium on Computer science education (SIGCSE '05)*. ACM, New York, NY, USA, 411-415.
- [7] Bond, T. G. 1976. BLOT: Bond's Logical Operations Test. Townsville, Australia. James Cook University.
- [8] Bond, T. G. Piaget and measurement I: The twain really do meet. *Archives de Psychologie*. 63, 245 (1995), 71-87.
- [9] Briggs, M. I. McCauley, M. H. Quenk, N. Hammer, A. 1998. *MBTI Handbook: A Guide to the development and use of the Myers-Briggs Type Indicator* Consulting (3rd edition). Psychologists Press.
- [10] Burney, G. M. 1974. *The construction and validation of an objective formal reasoning instrument*. Thesis. Division of Arts and Science., Uni. of Northern Colorado., Greeley., CO.
- [11] Byrne, J. H. Ph.D. 2014. Neuroscience Online. Department of Neurobiology and Anatomy, The UT Medical School at Houston. Retrieved September 4, 2014 from <http://neuroscience.uth.tmc.edu/>
- [12] Carlson, G. R. Streitberger, E. 1983. The Construction and Comparison of Three Related Tests of Formal Reasoning. *Science Education*. 67, 1 (1983), 133-140.
- [13] Carter, K. R. Ormrod, J. E. 1983. *The Social Sciences Piagetian Inventory (rev. edn.)*. Uni. of Northern Colorado, Dep. Ed. Psychology. Greeley, CO.
- [14] Crouch, C. H. Mazur, E. 2001. Peer instruction: Ten years of experience and results. *American Journal of Physics*. 69, 9 (2001), 970-977.
- [15] Digiampietri, A. L. 2012. Complementando o Aprendizado em Programação: Experiências no Curso de Sistemas de Informação da USP. In *VIII Simpósio Brasileiro de Sistemas de Informação* (São Paulo, SP, Brazil 2012). SBSI, 468-479.
- [16] Eckstein, S. G. and M. Shemesh. 1992. The rate of acquisition of formal operational schemata in adolescence: A secondary analysis. *Journal of Research in Science Teaching* 29, 5 (1992), 441-451.
- [17] Felder, R. M. Soloman, B. A. 2014. Learning styles and strategies. North Carolina State University. Retrieved September 4, 2014 from <http://www.ncsu.edu/felder-public/ILSdir/styles.htm>
- [18] Felder, R. M. Spurlin, J. 2005. Applications, reliability and validity of the index of learning styles. *International journal of engineering education* 21, 1 (2005), 103-112.
- [19] Furth, H. 1970. *An Inventory of Piaget's Developmental Tasks*. Thesis. Dept. of Psychology, Catholic Univ, Washington, DC.
- [20] Han, J. 2013. *Scientific Reasoning: Research, Development, and Assessment*. Doctoral Thesis. School of The Ohio State University. Columbus, OH.
- [21] Hmelo-Silver, C. E. 2004. Problem-based learning: What and how do students learn?. *Educational Psychology Review*. 16, 3 (2004), 235-266.
- [22] Hollander, M. Wolfe, D. A. Chicken, E. 2013. *Nonparametric statistical methods* (2nd. ed.). John Wiley & Sons, Danvers, MA.
- [23] Honzel, B. 2012. *The Effects of an Inquiry Based Data to Concept Curriculum*. Thesis. Montana State University.
- [24] Inhelder, B. Piaget, J. 1958. *The Growth of Logical Thinking From Childhood to Adolescence*. Basic Books Inc, New York, NY.
- [25] Ippolito, K. O. 2012. *Cognitive development and the attainment of critical thinking skills in associate degree nursing students*. Thesis. University of the Pacific Stockton, CA.
- [26] Kalat, J. W. 2008. Nature, Nurture, and Human Development. In *Introduction To Psychology* (8th. ed.). ch. 5, sec. 5.2, pp. 163 – 181. Thomson Wadsworth, Belmont, CA.
- [27] Kayes, D. C. 1999. Internal validity and reliability of Kolb's learning style inventory version 3. *Journal of Business and Psychology*. 20, 2 (Winter 2005), 249-257.
- [28] Kinnunen, P. Meisalo, V. Malmi, L. 2010. Have we missed something?: identifying missing types of research in computing education. In *Proceedings of the Sixth Int. Workshop on Computing education research* (New York, NY, USA 2010). ICER '10. ACM, New York, NY, 13-22.
- [29] Kitchenham, B. et al. 2007. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. Technical Report. School of Computer Science and Mathematics at Keele University. Department of Computer Science at University of Durham. Durham, UK.
- [30] Kolb, D. A. 1984. *Experiential learning: Experience as the source of learning and development*. (Vol. 1). Prentice-Hall, Englewood Cliffs, NJ.
- [31] Krathwohl, D. R. 2012. A revision of Bloom's taxonomy: An overview. *Theory Into Practice*. Routledge. 41, 4 (Aut. 2002), 212–218.
- [32] Kuhn, D. 2008. Formal Operations from a Twenty-First Century Perspective. *Human Dev.*, 51, 1(2008), 48–55.
- [33] Kurtz, B. L. 1980. Investigating the relationship between the development of abstract reasoning and performance in an introductory programming class. In *Proceedings of the eleventh SIGCSE technical symposium on Computer science education*. ACM. Kansas City, Missouri, USA. 110-117.
- [34] Lawson, A. 1978. The Development and Validation of a Classroom Test of Formal Reasoning. *Journal of Research in Science Teaching*. 15 (1978), 11 – 24.
- [35] Lawson, A. E. 2000. *Lawson Classroom Test of Scientific Reasoning*. Personal Communication. E-mail received from the author Feb. 2014.

- [36] Lawson, A. 2010. *Teaching Inquiry Science in Middle and Secondary Schools*. Sage, Publications, Inc. Thousand Oaks, CA.
- [37] Lawson, A. E., et al. 2000. What kinds of scientific concepts exist? Concept construction and intellectual development in college biology. *Journal of Research in Science Teaching*. 37, 9 (2000), 996-1018.
- [38] Lawson, A. E. *Keys to Lawson Classroom Test of Scientific Reasoning*, 2000 version. Personal Communication. E-mail received from the author Feb. 2014.
- [39] Lawson, A. E. 2004. The nature and development of scientific reasoning: A synthetic view. *Int. Journal of Science and Mathematics Education*. 2, 3 (2004), 307-338.
- [40] Lawson, Anton. E. 2003. *The neurological basis of learning, development and discovery: Implications for science and mathematics instruction*. Springer, Kluwer Academic Publishers, Dordrecht.
- [41] Longeot, F. 1966. L'échelle de développement de la pensée logique. BINOP, Paris.
- [42] Logan, R. A. 1991. *Relationships between the Arlin Test of Formal Reasoning and Classroom Multiple Choice*. Thesis. Emporia State University.
- [43] Lippman, J. 2012. *Improving and predicting novice reasoning about the evidentiary connection between studies and theories*. Doctoral Thesis. US, ProQuest Information & Learning. University of Illinois at Chicago, Chicago, IL.
- [44] Lunzer, E. A. 1965. Problems of Formal Reasoning in test situations. *Monographs of the Society for Research in Child Development*. 30, 2 (1965), 19 – 46.
- [45] Malmi, L. et al. 2014. Theoretical underpinnings of computing education research: what is the evidence?. In *Proceedings of the tenth annual conference on International computing education research*. (Glasgow, Scotland, Aug 2014). ICER '14. ACM, New York, NY, USA, 27-34.
- [46] Malmi, L. Sheard, J. Simon, Bednarik, R. Helminen, J. Korhonen, A. Myller, N. Sorva, J. Taherkhani, A. 2010. Characterizing research in computing education: a preliminary analysis of the literature. In *Proc. of the Sixth Int. workshop on Computing education research* (Aarhus, Denmark 2010). ICER '10. ACM, New York, NY, 3-12.
- [47] Manhães, M. B. L. et al. 2012. Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados: Uma Abordagem Quantitativa. In *VIII Simpósio Brasileiro de Sis. de Informação SBSI* (São Paulo, SP 2012), SBC, 468-479.
- [48] Mazur, E. 1997. *Peer Instruction: A User's Manual*. Prentice Hall, Upper Saddle River, NJ.
- [49] Mozzachiodi, R. Byrne, J. H. 2010. More than synaptic plasticity: role of nonsynaptic plasticity in learning and memory. *Trends in neurosciences*. 33, 1 (Jan. 2010), 17-26.
- [50] Parham, J. R. 2003. An assessment and evaluation of computer science education. *J. Comput. Sci. Coll.* 19, 2 (December 2003), 115-127.
- [51] Pears, A. Malmi, L. 2009. Values and objectives in computing education research. *ACM Trans. Comput. Educ.* 9, 3, Article 15 (September 2009), 6 pages.
- [52] Pears, A. Seidman, S. Malmi, L., et al. 2007. A survey of literature on the teaching of introductory programming. In *Working group reports on ITiCSE on Innovation and technology in computer science education* (2007). ITiCSE-WGR '07. ACM, New York, NY, USA, 204-223.
- [53] Piaget, J. 1957. *Logic and Psychology*. New York: Basic Books Inc, New York, NY.
- [54] Randolph, J. J. 2007. *Computer Science Education Research at the Crossroads: A Methodological Review of Computer Science Education Research: 2000-2005*. Doctoral Thesis. Utah State University. Logan, UT.
- [55] Raven, R. J. 1973. The Development of a Test of Piaget's Logical Operations. *Science Education*, 57,3 (Jul/Sep 1973) 377 – 385.
- [56] Roadrangka, V. 1985. *The construction and validation of the group assessment of logical thinking (GALT)*. Thesis, Uni. Of Georgia, Athens, GA.
- [57] Roberge, J. J. Flexer, B. K. 1982. The Formal Operational Reasoning Test. *The Journal of General Psychology*. 106, 1 (1982), 61-67.
- [58] Rowell, J. A. Hoffman, P. J. 1957. Group tests for distinguishing formal from concrete thinkers. *Journal of Research in Science Teaching*, 12 (1975), 157 – 164.
- [59] Russel, H. 1962. The Measurement of Reasoning Ability in Adolescents. *Ontario Journal of Educational Research*. 5, 1(1962), 33 – 40.
- [60] Shayer, M. Adey, P. Wylam, H. 1981. Group Tests of Cognitive Development Ideals and a Realization. *Journal of Research in Science Teaching*. 18, 2 (1981), 157 – 168.
- [61] Simons, D.J. Chabris, C.F. 2011. What people believe about how memory works: A representative survey of the US population. *PloS one*. 6,8, e22757 (Aug. 2011).
- [62] Sousa, V. D. Rojjanasrirat, W. 2011. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *Journal of Evaluation in Clinical Practice*. 17 (2011), 268–274.
- [63] Tisher, R. P. 1971. A Piagetian questionnaire applied to pupils in a secondary school. *Child Development*. 42, 5 (Nov 1971), 1633 – 1636.
- [64] Tobin, K. G. Capie, W. 1981. The Test of Logical Thinking (TOLT): The Development and Validation of a Group Test of Logical Thinking. *Educational and Psychological Measurement*, 41, 2 (Jul 1981), 413 – 423.
- [65] Tomlinson-Keasey, C. 1976. Can we develop abstract thought in college freshmen? *84th Annual American Psychological Association Conv.* (1976). Washington, D.C.
- [66] Veenman, M. V. J. Van Hout-Wolters, B. H. A. M. Afflerbach, P. 2006. Metacognition and learning: Conceptual and methodological considerations. *Metacognition and learning*. 1, 1 (Apr. 2006), 3–14.

Uma Ferramenta para Busca Não Estruturada em Código AspectJ

Alternative Title: A Tool for Searching in Unstructured Code AspctJ

Rodrigo Castro Gil
Universidade Federal de
Santa Maria
Avenida Roraima, 100
Santa Maria/RS, Brasil
rcgil@inf.ufsm.br

Eduardo Kessler Piveta
Universidade Federal de
Santa Maria
Avenida Roraima, 100
Santa Maria/RS, Brasil
piveta@inf.ufsm.br

Deise de Brum Saccol
Universidade Federal de
Santa Maria
Avenida Roraima, 100
Santa Maria/RS, Brasil
deise@inf.ufsm.br

Cristiano de Faveri
Universidade Nova de Lisboa
Caparica, Portugal
c.faveri@campus.fct.unl.pt

RESUMO

Com o aumento do tamanho dos sistemas de informação e de sua complexidade e com as limitações que as IDEs disponíveis atualmente possuem em relação a buscas mais complexas em código fonte, ferramentas que possam auxiliar desenvolvedores de software na recuperação de informações relevantes tornam-se muito úteis. Neste contexto, este artigo apresenta uma ferramenta que viabiliza a recuperação de informações em código AspectJ de forma não estruturada. Essa ferramenta permite a busca nas estruturas utilizando apenas o valor sintático da consulta ou agregando valor semântico e com isso melhorando seus resultados.

Palavras-Chave

Recuperação de Informação, Código Fonte, AspectJ

ABSTRACT

With the increasing size of information systems and their complexity and the limitations that currently have available the IDEs for more complex searches in source code, tools that can help software developers in the retrieval of relevant information becomes very useful. In this context, this paper presents a tool that enables the retrieval, informations in AspectJ code in an unstructured way. This tool allows you to search the structures using only the syntactic value of the query or adding semantic value and thereby improving their results.

Categories and Subject Descriptors

D.3.0 [Programming Languages]: General—Standards

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015

General Terms

Languages

Keywords

Information Retrieval, Code Sources, AspectJ

1. INTRODUÇÃO

No âmbito de recuperação de informação (RI), a recuperação de dados significa, principalmente, determinar quais documentos de uma coleção têm as palavras-chave pesquisadas em um determinado banco de dados, o que muitas vezes pode não satisfazer, de modo completo, a consulta realizada pelo usuário. Já a recuperação de informação é a interpretação dos conteúdos de informações presentes nos documentos de uma coleção e os seleciona de acordo com um grau de relevância. Esta interpretação envolve extrair, por meio de processamento de linguagem natural, informações do documento de texto [8].

A principal diferença entre a recuperação de informação e a de dados fundamenta-se no fato de que a primeira abrange textos de linguagem natural que muitas vezes não são bem estruturados e podem apresentar ambiguidade; já a segunda, lida com dados caracterizados pela estrutura e semântica bem definidas [1].

Os sistemas informatizados crescem de tamanho e de complexidade a cada dia. A busca pela qualidade em um sistema de software envolve, não somente práticas estabelecidas em um processo de desenvolvimento, mas também a análise das estruturas de código (análise estática).

No contexto de busca em código fonte existem ferramentas que realizam a análise estática, as quais operam basicamente através da obtenção de dados de um programa, da execução de algoritmos de análise e da apresentação dos resultados. A obtenção de dados de um programa é um ponto crítico para essas ferramentas, pois de acordo com o seu propósito, diferentes visões podem ser necessárias [7].

A análise estática visa a diversos objetivos, tais como descobrir oportunidades de refatoração em código, buscar falhas de segurança, assegurar convenções de escrita, etc. Em último caso, a busca e a recuperação também se aplicam a propósitos mais simples, como encontrar determinado elemento no código fonte a fim de compreender suas estruturas. A recuperação de informações de um programa é um ponto crítico na análise estática. As principais IDEs podem ser ineficientes para buscas mais complexas. Como, por exemplo, buscar todas as classes que sejam públicas e tenham um método que retorne um inteiro.

Diante da necessidade de busca e de recuperação de informação em código fonte, este artigo apresenta uma aplicação para busca textual, utilizando técnicas de busca não estruturada, em código fonte com suporte para programas orientados a objetos e orientados a aspectos. O objetivo é possibilitar que os programadores possam refinar as consultas nos repositórios de código, dando sentido semântico às palavras reservadas da linguagem, possibilitando assim definir contextos nas consultas. Esses contextos são definidos através de regras opcionais que ajudam a refinar as consultas.

Esse artigo está estruturado da seguinte forma: a Seção 2 mostra informações sobre as tecnologias utilizadas, a Seção 3 mostra a visão geral da abordagem assim como a forma de marcação dos arquivos, a indexação dos mesmos e apresenta algumas regras definidas com o objetivo de refinar as consultas, a Seção 4 apresenta um estudo de caso e seus resultados, a Seção 5 apresenta um breve comparativo com alguns trabalhos relacionados e a Seção 6 apresenta as conclusões do trabalho.

2. TECNOLOGIAS ASSOCIADAS

Para a implementação da aplicação foram utilizadas duas tecnologias associadas: Lucene e AOPJungle, as quais são apresentadas a seguir.

2.1 Lucene

O Apache Lucene [10] é uma biblioteca de software livre para indexação e recuperação de informações, originalmente escrita em Java. As APIs do Lucene focam principalmente na indexação e na procura de texto. Elas podem ser usadas para criar recursos de procura para aplicativos, como clientes de e-mail, listas de correspondência, procuras na Web, procuras em banco de dados, etc. Web sites como Wikipedia¹, TheServerSide², jGuru³ e LinkedIn⁴ foram desenvolvidos utilizando como motor de busca o Lucene [11].

As APIs do Lucene focam principalmente na indexação e na procura de texto. Elas podem ser usadas para criar recursos de procura para aplicativos, como clientes de e-mail, listas de correspondência, procuras na Web, procuras em banco de dados, etc. Embora tenha sido implementado em Java, devido à sua popularidade e à sua comunidade de desenvolvedores atualmente o Lucene conta com portabilidade e integração com várias outras linguagens (C/C++, C#, Ruby, Perl, Python, PHP, entre outras).

¹www.wikipedia.org

²www.theserverside.com

³www.jguru.com

⁴www.linkedin.com

Dentre os recursos fornecidos pelo Lucene, os seguintes se destacam:

- retorna os documentos classificados como relevantes a partir do cálculo de uma pontuação atribuída a eles.
- possibilita vários tipos de consultas, como consultar uma frase específica (*PhraseQuery*), utilizar um ou mais caracteres como coringa (*WildcardQuery*), realizar a busca em um intervalo específico (*RangeQuery*), buscar por semelhança (*FuzzyQuery*), entre outros;
- admite a análise de expressões de consultas completas digitadas pelo usuário;
- usa um mecanismo de bloqueio baseado em arquivo para impedir modificações de índices simultâneos;
- permite a procura e a indexação simultaneamente.

2.2 AspectJ

A orientação a aspectos (OA) é um paradigma que estende a orientação a objetos (e outros, como o paradigma estruturado) introduzindo novas abstrações. Estes novos elementos são destinados a suprir deficiências na capacidade de representação de algumas situações [12].

Um dos elementos centrais da OA é o conceito de interesse, que são as características relevantes de uma aplicação. A OA introduz novas abstrações de modularização e mecanismos de composição para melhorar a separação de interesses transversais⁵. Essas abstrações são: os aspectos (*aspects*), pontos de junção (*joinpoints*), pontos de corte (*pointcuts*), adendos (*advices*) e declaração intertipos *inter-type declarations*. A seguir é mostrada uma breve definição dessas construções.

Aspectos

Aspecto é o termo usado para denotar a abstração da OA que dá suporte a um melhor isolamento de interesses transversais. Em outras palavras, um aspecto corresponde a um interesse transversal e constitui uma unidade modular projetada para entrecortar um conjunto de classes e objetos do sistema.

Pontos de junção

Um ponto de junção é um ponto bem específico da execução de um sistema. Alguns exemplos de pontos de junção são: chamadas a métodos, execuções de métodos, leituras de atributos, modificações de atributos, etc. Através do conceito de pontos de junção, torna-se possível especificar o relacionamento entre aspectos e classes.

Pontos de corte

São os pontos responsáveis por detectarem quais pontos de junção o aspecto deverá interceptar.

⁵São chamados interesses transversais aqueles que não podem ser modularizados em classes.

Adendos

Adendo é um construtor especial semelhante a um método de uma classe, que define o comportamento dinâmico executado quando são alcançados um ou mais conjuntos de junção definidos previamente. Em outras palavras, adendos são recursos de entrecorte transversal que afetam o comportamento dinâmico de classes e objetos.

Declarações intertipos

As declarações intertipos especificam novos atributos e/ou métodos nas classes que um aspecto entrecorta, ou modificam o relacionamento entre classes e entre classes e interfaces. Ao contrário de adendos, que operam de forma dinâmica, as declarações intertipos operam estaticamente, em tempo de compilação. Além de especificarem elementos que entrecortam classes de um sistema, aspectos podem possuir métodos e atributos internos, como classes OO.

AspectJ estende Java com suporte para dois tipos de interesses transversais, o dinâmico e o estático. O primeiro torna possível definir qual execução adicional deve se executar em um determinado ponto durante a execução do programa. O segundo torna possível definir novas operações em tipos existentes.

2.3 AOPJungle

O AOPJungle [7] é um *framework* desenvolvido para simplificar o trabalho de extração e cálculo de fragmentos de programas orientados a aspectos, organizando os elementos estruturais, seus relacionamentos e partes do código em um modelo, o qual pode ser estendido para finalidades específicas. Além disso, esse *framework* disponibiliza algumas informações relevantes em relação ao programa, tais como o número de linhas de um método, de um *advice*, de um aspecto, o número de elementos descendentes de um aspecto, a profundidade da árvore de herança, etc., sendo útil a outras finalidades tais como atividades de análise estática e de extração de métricas.

O *framework* fornece uma API para acesso à informações de um programa e permite ainda consulta através da OQL (*Object Query Language*) [4], a fim de filtrar fragmentos de interesse do código. O AOPJungle foi implementado em AspectJ [9] para coleta de informações de programas escritos em AspectJ, e utiliza APIs da plataforma Eclipse para extração das informações acerca do programa. A Figura 1 ilustra a arquitetura geral do *framework* AOPJungle.

O módulo de metamodelo representa o programa em uma estrutura orientada a objetos na qual a API busca informações solicitadas pelos clientes do *framework*. O módulo de extração é responsável por ler e por transformar os elementos do programa no metamodelo correspondente.

À medida que os elementos estruturais são carregados, suas dependências estáticas são resolvidas pelo analisador de dependência. É esse módulo, por exemplo, que resolve quais pontos do programa um *advice* afeta. O processador de incrementos é responsável por computar determinados cálculos em relação ao programa e deixá-los disponíveis para consulta, como por exemplo, o número de linhas de cada método, *advice*, classe ou aspecto. Uma característica importante do AOPJungle é seu suporte de metamodelo a mul-

tipojetos, o qual permite consultas considerando diversos sistemas ao mesmo tempo.

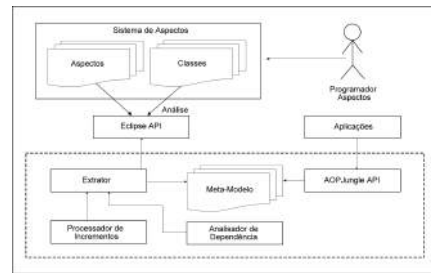


Figura1: Arquitetura do AOPJungle [7]

3. UMA FERRAMENTA PARA BUSCA NÃO ESTRUTURADA

A Figura 2 mostra a arquitetura da ferramenta proposta. A ferramenta é composta pelo *Index Engine* e pelo *Search Engine*. O *Index Engine* realiza a indexação dos arquivos, a qual se divide em extração de código (feita pelo AOPJungle), anotação de código (feita pela ferramenta de busca), e indexação do código anotado implementada utilizando o Lucene. Logo após, é mostrado onde os índices criados são mantidos. O *Search Engine* realiza pesquisa nos índices arquivados, a qual se divide em uma interface onde o usuário realiza a sua consulta, um analisador e transformador dessa consulta, um executor da consulta e um visualizador dos resultados.

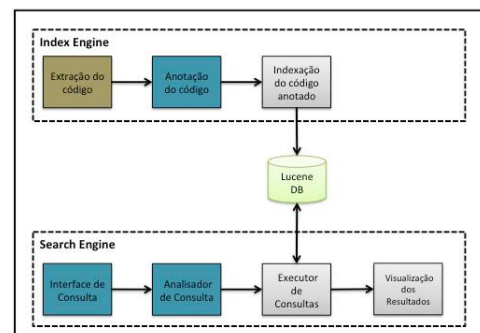


Figura2: Arquitetura da Aplicação

Buscas não estruturadas, ou seja, sem um formato, regra ou sequência padronizados, dependem dos critérios de classificação dos algoritmos usados pelos motores de busca aos quais elas são submetidas. O usuário final não pode mudar a maneira como um algoritmo classifica as palavras ou os arquivos, mas ele pode interagir com o sistema formulando ou reformulando uma consulta [6]. Essa interação é uma parte crucial para o processo de recuperação de informação, e pode determinar se o motor de busca está realizando um serviço eficaz.

Os clientes desta aplicação em geral saberão quais informações são relevantes nos resultados. Para melhorar esses resultados, é proposto dar sentido semântico aos termos da busca, utilizando palavras reservadas para isso, como por exemplo *public*, *String*, *return*, *int*, etc.

3.1 Indexação

O AOPJungle realiza a extração dos dados contidos nos documentos e toda vez que for encontrada uma palavra reservada será incluída a marcação definida. Por exemplo, ao encontrar o termo “public”, ele será identificado como um modificador e se essa palavra for parte da definição de uma classe esta receberá a marcação “«classModifier»PUBLIC”. O mesmo acontecerá com os métodos, aspectos etc.

Com as informações importantes retiradas e marcadas é criado um metamodelo, o qual é indexado. Esse metamodelo permite a construção de consultas com sentido semântico, facilitando, ao autor das consultas, fornecer informações mais precisas acerca das suas necessidades, diminuindo assim os falsos positivos, ou seja arquivos retornados após uma consulta que não satisfaçam corretamente as necessidades de informação dessas consultas.

Para uma melhor identificação dos arquivos com as informações retiradas dos arquivos originais, esse metamodelo foi definido com a extensão “.idx”, apenas para o processo de análise da aplicação.

A seguir é mostrado um exemplo de uma classe Java com alguns atributos, métodos de leitura e de escrita, e logo após o modelo marcado gerado a partir dessa classe.

```
public class Aluno{
    public Integer matricula;
    public String nome;
    public String curso;

    public String getCurso() {
        return curso;
    }
    public void setCurso(String curso) {
        this.curso = curso;
    }
    public Integer getMatricula() {
        return matricula;
    }
    public void setMatricula(Integer matricula) {
        this.matricula = matricula;
    }
    public String getNome() {
        return nome;
    }
    public void setNome(String nome) {
        this.nome = nome;
    }
}
```

Arquivo Marcado:

```
«packageName»aluno«className»Aluno
«classModifier»PUBLIC«attributeName»matricula
«attributeName»nome«attributeName»curso
«attributeType»Integer«attributeType»String
«attributeType»String«attributeModifier»PUBLIC
«attributeModifier»PUBLIC«attributeModifier»PUBLIC
«methodName»getCurso«methodName»setCurso
«methodName»getMatricula«methodName»setMatricula
«methodName»getNome«methodName»setNome
«returnType»String«returnType»void«returnType»Integer
«returnType»void«returnType»String«returnType»void
«methodModifier»PUBLIC«methodModifier»PUBLIC
```

```
«methodModifier»PUBLIC«methodModifier»PUBLIC
«methodModifier»PUBLIC«methodModifier»PUBLIC
«parameterName»curso«parameterName»matricula
«parameterName»nome«parameterType»String
«parameterType»Integer«parameterType»String
```

A Figura 3 mostra o resultado da indexação de um repositório onde se encontra a classe mostrada como exemplo, sendo possível então, recuperar as informações presentes nos documentos.

Figura3: Saída Indexação

3.2 Transformação das Consultas

Com o objetivo de melhorar a precisão das buscas realizadas nos repositórios, foram adotadas algumas regras para a realização das consultas. Para facilitar a transformação das consultas e oferecer aos usuários a possibilidade de refinar suas buscas melhorando assim os resultados, a seguir são listadas essas regras.

Regra 1: Para dar sentido semântico às consultas elas devem iniciar com o símbolo “::”, o qual define que a próxima palavra definirá o contexto da consulta.

Se a consulta não for iniciada pelo símbolo proposto, a busca será realizada levando em conta apenas o valor textual, não observando o valor semântico das palavras reservadas. Por exemplo, se o usuário digitar apenas “String”, a consulta retornará todas a ocorrências dessa palavra, não importando se ela define um tipo de retorno ou um tipo de um atributo, etc.

Regra 2: Para realizar uma consulta que retorne o nome de um pacote, o seguinte texto é necessário: “::package nome do pacote”, onde “package” define que a busca será realizada apenas nos pacotes presentes no repositório indexado. Por exemplo, para buscar um pacote com o nome “aluno”, o processo seria esse:

Consulta do Usuário: ::package aluno

Consulta traduzida: «packageName»aluno

Regra 3: Para buscar as informações de uma classe a consulta deve ser iniciada por “::class”, seguida das informações as quais pretende-se encontrar. Por exemplo, para buscar uma classe que tenha o nome “Aluno”, as consultas seriam essas:

Consulta do usuário: ::class Aluno

Consulta traduzida: «className»Aluno

Uma observação importante, se for pretendido buscar uma classe utilizando o nome da classe, o nome deve vir logo após a definição “::class”, mas se a procura pretendida não precisar do nome da classe, esse não precisa constar na consulta. A seguir um exemplo de uma consulta feita com base no modificador da classe.

Consulta do usuário: **::class public**

Consulta traduzida: **«classModifier»public**

Regra 4: Para recuperar as informações acerca de um interface a consulta deve ser iniciada por “::interface”, seguida das informações as quais pretende-se encontrar. Por exemplo, para buscar uma interface que tenha o nome “MinhaInterface”, as consultas seriam essas:

Consulta do usuário: **::interface MinhaInterface**

Consulta traduzida: **«interfaceName»MinhaInterface**

Da mesma forma que nas classes não é necessário que a consulta tenha como termo o nome da interface, mas se ele for utilizado deve vir logo após da definição “::interface”.

Regra 5: Para procurar informações acerca de um método a consulta deve ser iniciada por “::method”, completada das informações as quais serão buscadas. A seguir um exemplo de uma consulta que busca todos os métodos privados.

Consulta do usuário: **::method private**

Consulta transformada: **«methodModifier»private**

Regra 6: Se a busca pretendida for acerca do retorno de um método ou de um *advice around*, a mesma deve ser feita diretamente nos tipos de retornos, e deve ser iniciada por “::return”. Aqui um exemplo de uma pesquisa feita para encontrar todos os métodos, ou *advice around*, que tenham retorno do tipo *String*:

Consulta do usuário: **::return String**

Consulta transformada: **«returnType»String**

Regra 7: Para consultar informações sobre parâmetros de um método a consulta deve ser iniciada por “::parameter”, se for pretendido buscar através do nome do parâmetro, o nome deve ser o primeiro item da consulta, assim como nas classes e métodos. A seguir o exemplo de uma consulta para buscar todos os parâmetros com o nome “args”.

Consulta do usuário: **::parameter args**

Consulta transformada: **«parameterName»args**

Regra 8: Para consultar informações acerca dos atributos (*fields*) a consulta deve iniciar com “::field”. A seguir um exemplo de uma consulta que retorna todos os atributos privados.

Consulta do usuário: **::field private**

Consulta transformada: **«attributeModifier»private**

Regra 9: Para consultar informações acerca de um aspecto a consulta deve ser iniciada por “::aspect” seguida das informações desejadas. A seguir um exemplo de uma consulta que busca todos os aspectos privados:

Consulta do usuário: **::aspect private**

Consulta transformada: **«aspectModifier»private**

Regra 10: Para consultar acerca de um *advice*, seja um *adviceAfter*, um *adviceAfterReturning*, um *adviceAfterThrowing*, um *adviceBefore* ou um *adviceAround*, é necessário

buscar direto pelo *kind* que define a forma do *advice*. Por exemplo, uma consulta que retorna todos os *advice before* deve ter a seguinte estrutura:

Consulta do usuário: **::adviceKind Before**

Consulta transformada: **«adviceKind»Before**

Regra 11: Para consultar acerca de parâmetros de um *advice*, a consulta deve ser iniciada por “::aspectParameterType”, se a consulta tomar como base o tipo do parâmetro, e se a consulta tomar como base o nome do parâmetro esta deve ser iniciada por “::aspectParameterName”. A seguir um exemplo de cada consulta:

Exemplo 1:

Consulta do usuário: **::aspectParameterType Graphics**

Consulta transformada: **«aspectParameterType»Graphics**

Exemplo 2:

Consulta do usuário: **::aspectParameterName gr**

Consulta transformada: **«aspectParameterName»gr**

Regra 12: Também é possível construir consultas mais elaboradas utilizando os operadores lógicos “AND” e “OR”. Para isso basta o usuário inserir o “AND” ou o “OR” entre os termos da consulta. Se o separador dos termos for apenas um espaço em branco o Lucene assume ser um “OR”. Lembrando que para a troca de contexto é necessário que o usuário insira o símbolo “::”, para que o analisador das consultas reconheça essa troca. Por exemplo, se a primeira parte da consulta for sobre uma classe e a segunda for sobre um método para marcar a troca de contexto é preciso inserir “::method”, para que o analisador troque de contexto. A seguir é mostrado um exemplo utilizando o operador “AND”.

Exemplo 1: Consulta para buscar os métodos que tenham o modificador *public* e sejam da classe “Aluno”.

Consulta do usuário: **::method public AND ::class Aluno**

Consulta transformada: **«methodModifier»public AND «className»Aluno**

Exemplo 2 : Uma busca que retorne todos métodos com o modificador *public* ou todos os métodos que tenham o retorno do tipo *String*

Consulta do usuário: **::method public ::return String**

Consulta transformada: **«methodModifier»public OR «returnType»String**

Regra 13: Outro recurso do Lucene que é explorado para refinar as consultas, é a utilização dos operadores unários de adição e subtração. A seguir é mostrada a aplicação desses operadores.

Para buscar os documentos que tenham um aspecto com o nome “NextBlock” e não possam ser públicos basta entrar com a seguinte consulta: **::aspect NextBlock - public**, pois o operador de subtração exclui todos os documentos que tenham o termo que vem logo após ele.

Da mesma forma é possível usar o operador de adição para buscar todos os arquivos que tenham obrigatoriamente um aspecto público e que possam ter o nome de “NextBlock”. Para tal basta entrar com a seguinte consulta: **::aspect NextBlock + public**, pois o operador de adição exige que os documentos, para serem considerados relevantes, tenham o termo que vem logo após ele. Por convenção da aplicação,

esses operadores deverão sempre ser usados antes do segundo termo da consulta.

Todos os itens de uma consulta recebem os caracteres barra e aspas no início e no fim. Isso é necessário para que o Lucene processe corretamente os caracteres *guillemets* inseridos para marcar cada item, dessa forma o Lucene reconhece esses caracteres e retorna os resultados com precisão.

As regras definidas nessa seção não são exaustivas para todas as possibilidades de consultas a serem feitas. Elas abrangem as estruturas mais utilizadas da linguagem, e foram criadas com a preocupação de não deixar as consultas com uma estrutura rígida. A Tabela 1 mostra de forma resumida as regras definidas.

Tabela1: Regras para Consultas

Regra	Função	Sintaxe
1	Definir todos os contextos	::
2	Contexto pacotes	::package
3	Contexto classes	::class
4	Contexto interfaces	::interface
5	Contexto métodos	::method
6	Contexto retornos	::return
7	Contexto parâmetros métodos	::parameter
8	Contexto atributos	::field
9	Contexto aspectos	::aspect
10	Contexto advices	::adviceKind
11	Contexto parâmetros advices	::asp
12	AND e OR	AND, OR ou um espaço em branco
13	Exigência Ou negação	+ para exigência, e - para exclusão, sempre antes do segundo termo da consulta

4. ESTUDO DE CASO

Para demonstração do funcionamento da aplicação foi utilizado, como repositório, um *workspace* com alguns projetos Java e alguns projetos AspectJ, o qual teve a sua indexação mostrada na Figura 3. A maior parte das consultas foram realizadas no projeto AOPTetris. Esse projeto foi escolhido por ser implementado em AspectJ e possuir aspectos e classes possibilitando assim consultas mais elaboradas.

A Figura 4 mostra o resultado da consulta “::aspect NextBlock”, a qual retornou o arquivo: *NextBlock.idx*, do pacote *logic*, sendo esse o único arquivo que satisfaz essa consulta entre os arquivos indexados.

```
Encontrados 1 documento(s) (em 91milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNextBlock.idx
```

Figura4: Consulta ::aspect NextBlock

A Figura 5 mostra o retorno da consulta “::aspect NextBlock public”, a qual retorna o arquivo *NextBlock.idx* que satisfaz

o primeiro elemento da consulta, por ter um aspecto com o nome *NextBloc*, e todos os outros arquivos satisfazem o segundo elemento dessa consulta, por terem aspectos públicos.

```
Encontrados 10 documento(s) (em 38milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNextBlock.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.developmentDesignCheck.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.guiGameInfo2.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.coreAspectTest.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.guiGameInfo.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.TestAspect2.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.highscoreCounter.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.guiMenu.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.highscoreLevels.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNewBlocks.idx
```

Figura5: Consulta ::aspect NextBlock public

A Figura 6 mostra o resultado de uma consulta onde cada um dos termos da consulta está em um contexto. O primeiro termo busca pelo nome de um aspecto e o segundo por um ponto de corte do tipo *call*. O primeiro arquivo retornado satisfaz o primeiro termo da consulta e os outros três arquivos correspondem ao retorno do segundo termo. Essa consulta é construída da seguinte forma: “::aspect NextBlock ::pointcut call”.

Como foi abordado na regra 3.2, tanto na consulta da Figura 5 quando na consulta da Figura 6, o que separa os termos da consulta é um espaço em branco, o qual é interpretado como um “OR”, portanto os arquivos precisam satisfazer apenas um dos termos da consulta para serem considerados como relevantes.

```
Encontrados 4 documento(s) (em 33milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNextBlock.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.highscoreCounter.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNewBlocks.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.highscoreLevels.idx
```

Figura6: Consulta ::aspect NextBlock ::pointcut call

Uma consulta com os mesmos termos da consulta anterior, mas que os arquivos precisam satisfazer os dois termos da consulta para serem considerados relevantes, é construída da seguinte forma, “::aspect NextBlock + ::pointcut call”. E tem seu retorno mostrado na Figura 7, e apesar de ter o mesmo retorno da consulta mostrada na figura 4.

```
Encontrados 1 documento(s) (em 118milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNextBlock.idx
```

Figura7: Consulta ::aspect NextBlock + ::pointcut call

O retorno mostrado na Figura 8 refere-se à seguinte consulta: “::adviceKind Around + ::returnType int”. Essa consulta retorna todos os arquivos que tenham um *advice Around* e um retorno do tipo *int*. O resultado dessa consulta não garante é que esse retorno seja desse *advice*. O retorno poderá ser de um método do mesmo aspecto que faz parte do mesmo arquivo.

```
Encontrados 2 documento(s) (em 39milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNewBlocks.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNextBlock.idx
```

Figura8: Consulta ::adviceKind Around + ::returnType int

A Figura 9 mostra o resultado da seguinte consulta: “(::aspect NextBlock + ::adviceKind After) ::class GameInfo”.

Ela mostra que é possível utilizar os operadores “AND”, e “OR”, com resultados combinados possibilitando assim um maior refinamento de consultas. No exemplo citado, para um documento ser considerado relevante, ele precisa satisfazer as duas condições do primeiro termo da consulta ou satisfazer o segundo termo dela. Também é mostrado no exemplo que é possível trocar de contexto inúmeras vezes, na consulta mostrada o contexto começa em um aspecto, muda para um adendo, e depois muda para uma classe.

```
Encontrados 2 documento(s) (em 142milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNextBlock.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.guiGameInfo.idx
```

Figura9: (Consulta `::aspect NextBlock + ::advice-Kind After`) `::class GameInfo`

Como a coleção de documentos indexados é conhecida, conforme mostrado na figura 3, foi possível aplicar as duas técnicas de medição de desempenho em recuperação de informação, mais utilizadas, que são:

- Precisão (P), que é a fração de documentos retornados que é relevante;
- Revocação (R), que é a fração dos documentos relevantes retornados.

Para uma melhor compreensão da aplicação dessas técnicas o conjunto de arquivos indexados, foi dividido em quatro subconjuntos como mostra a Figura 10. Com base nessa divisão é possível definir a Precisão por : $P = \frac{A}{A+B}$, e a Revocação por: $R = \frac{A}{A+D}$.

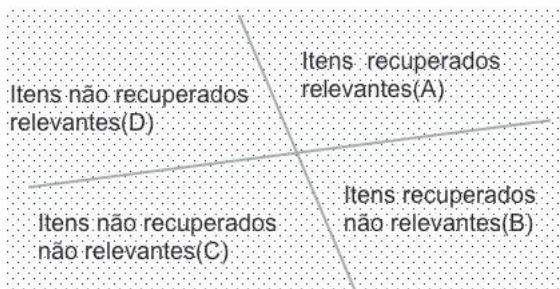


Figura10: Divisão do conjunto de arquivos

A Figura 11, mostra o resultado de uma busca por: “GameInfo”. O resultado mostrado é obtido através da busca com base apenas no valor textual da palavra. Neste caso o resultado é o mesmo da consulta que agrega valor semântico, assumindo-se que o usuário procura pela classe “GameInfo”, realizando a seguinte consulta: “`::class GameInfo`”, com o mesmo resultado a Precisão e a *Recall* têm o mesmo valor nas duas consultas, como é mostrado na Tabela 2.

```
Encontrados 1 documento(s) (em 49milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.guiGameInfo.idx
```

Figura11: Consulta GameInfo

A Figura 12 mostra o resultado da consulta pelo termo “int”. Como é possível observar a consulta teve como resultado

quatorze documentos, assumindo que o usuário procura pelos documentos que tenham atributos do tipo *int*, esta pode ser melhorada para: “`::field int`”, a qual tem seu resultado mostrado na Figura 13, e apresenta doze arquivos como retorno. Num primeiro olhar, essa diferença pode não parecer significativa, mas como é mostrado na Tabela 2 a Precisão tem uma redução em quase 15%.

```
Encontrados 14 documento(s) (em 27milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNewBlocks.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.logicBlocks.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.eventInterfaceEventListener.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.guiBlockPanel.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.highscoreLevels.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNextBlock.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNewBlocks.idx
/Users/Gil/workspace/indexer/meta_modelo/testalampadaLampada.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.coreAspectTetris.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.guiGameInfo.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.highscoreCounter.idx
```

Figura12: Consulta int

```
Encontrados 12 documento(s) (em 45milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.eventInterfaceEventListener.idx
/Users/Gil/workspace/indexer/meta_modelo/testalampadaLampada.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.logicBlocks.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.guiGameInfo.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.guiBlockPanel.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.highscoreCounter.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.coreAspectTetris.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.highscoreLevels.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNextBlock.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNewBlocks.idx
```

Figura13: Consulta `::field int`

A diferença de precisão é maior ainda, quando realizada uma busca apenas pelo termo “String”, a qual tem seu resultado mostrado na Figura 14. Assumindo-se que o usuário busca por todos os arquivos que possuam um retorno do tipo *String*, essa consulta pode ser melhorada para: “`::return String`”, e tem seu resultado mostrado na Figura 15. Na Tabela 2 é possível constatar que a Precisão, entre as consultas, tem uma variação de mais de 70%, sendo possível constatar que o refinamento de consulta proposto tem resultados positivos.

```
Encontrados 11 documento(s) (em 19milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspectsTestAspect.idx
/Users/Gil/workspace/indexer/meta_modelo/alunoAluno.idx
/Users/Gil/workspace/indexer/meta_modelo/testalampadaTestalampada.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.logicTetrisImages.idx
/Users/Gil/workspace/indexer/meta_modelo/indexerIndexer.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNewBlocks.idx
/Users/Gil/workspace/indexer/meta_modelo/desenhoDesenhoApp.idx
/Users/Gil/workspace/indexer/meta_modelo/testalampadaLampada.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.guiBlockPanel.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.coreAspectTetris.idx
```

Figura14: Consulta String

```
Encontrados 3 documento(s) (em 34milliseconds)
/Users/Gil/workspace/indexer/meta_modelo/alunoAluno.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.aspects.logicNewBlocks.idx
/Users/Gil/workspace/indexer/meta_modelo/com.ajtetris.logicBlocks.idx
```

Figura15: Consulta `::return String`

Tabela2: Comparação da Precisão e da Revocação entre as consultas

Consulta	Revocação	Precisão
“GameInfo”	$R = \frac{1}{1+0} = 1$	$P = \frac{1}{1+0} = 1$
“ <code>::class GameInfo</code> ”	$R = \frac{1}{1+0} = 1$	$P = \frac{1}{1+0} = 1$
“int”	$R = \frac{12}{1+0} = 1$	$P = \frac{12}{12+2} = 0,8571$
“ <code>::field int</code> ”	$R = \frac{1}{1+0} = 1$	$P = \frac{12}{12+0} = 1$
“String”	$R = \frac{3}{1+0} = 1$	$P = \frac{3}{3+8} = 0,2727$
“ <code>::return String</code> ”	$R = \frac{3}{1+0} = 1$	$P = \frac{3}{3+0} = 1$

Como pode ser constatado na Tabela 2 a revocação se manteve igual, em todas as consultas comparadas, isso acontece pelo fato de que nenhuma das consultas deixou de retornar algum arquivo relevante.

5. TRABALHOS RELACIONADOS

NerdyData [3] é um motor de busca com a finalidade de pesquisar o código de *websites*. Com a capacidade de indexar HTML, JavaScript, CSS e texto simples, o Nerdyydata permite uma série de consultas como: uma forma livre de busca para uma determinada frase. Também é possível realizar uma pesquisa comparativa de até cinco termos para encontrar domínios, entre outras. A ferramenta apresentada neste artigo tem como finalidade pesquisar em grandes repositórios de código fonte enquanto o NerdyyData tem como foco repositórios de sites web.

Krugle é um portal de busca de código aberto que em repositórios como o Apache, JavaDocs e SourceForge entre outros [5]. Você pode pesquisar por código em C ++, Java, Perl, Python, SQL, Ruby, XML, HTML, etc. Krugle tem um recurso de busca avançada que pode ajudar a diminuir a APIs certas, bibliotecas, código de amostra ou documentação. A partir da página de resultados, você pode navegar para o projeto desenvolvido com o código. O Krugle também tem suporte para AspectJ porém ele busca pelo valor sintático das palavras enquanto a aplicação apresentada permite agregar valor semântico às consultas.

Ohloh Code é apresentado como um dos maiores e mais abrangentes motores de busca de código com mais de 10 bilhões de linhas de código indexados e atualizados nos diretórios de software FOSS [2]. Ele indexa todos os arquivos de texto para pesquisa e tem destaque de sintaxe pois suporta 43 linguagens de programação. A sintaxe de consulta de pesquisa possui flexibilidade para procurar diferentes classes de código. O Ohloh Code atualmente não tem suporte a expressões regulares, enquanto a proposta apresentada possui.

6. CONCLUSÃO

Esse artigo apresentou uma ferramenta que auxilia o trabalho dos programadores, possibilitando o refinamento de consultas em código fonte. Para desenvolver a aplicação com suporte à linguagem AspectJ, foi necessário um estudo para um melhor conhecimento das estruturas da linguagem. Em seguida foi feita uma investigação acerca do funcionamento do *framework* AOPjungle para entender como é feita a extração das informações presentes no código sendo possível então realizar as marcações necessárias e o refinamento das consultas.

As consultas, para que seja possível serem refinadas, apresentam uma estrutura mais definida o que foge um pouco da proposta do trabalho. Por essa razão será desenvolvido um processo para que seja possível apresentar os resultados com base na localização dos objetos das consultas deixando assim mais livre a estrutura das consultas.

A aplicação proposta foi desenvolvida com suporte para as linguagens de programação AspectJ e Java, mas pode ser estendida para outras linguagens de programação. Com a utilização dessa ferramenta os usuários podem construir

consultas com complexidade maior obtendo resultados, com base nos testes realizados na Seção 4, com boa precisão. Vale lembrar que essa aplicação tem como usuário final programadores, ou outros programas, pois ela depende do *feedback* desse usuário para que se possa ter certeza se as necessidades de informação pretendidas foram atendidas.

Essa aplicação não tem como objetivo classificar os documentos retornados por ordem de relevância. Isso é feito pelo Lucene o qual usa um algoritmo de classificação próprio para esse propósito. A classificação com base nas prioridades da linguagem é um ponto importante principalmente para repositórios muito grandes, por isso os autores já estão trabalhando em pesquisas para oferecer essa classificação. Também será feita a integração com as IDEs para a visualização dos resultados.

7. REFERÊNCIAS

- [1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
- [2] S. Basu. Ohloh.
<http://www.makeuseof.com/tag/open-source-matters-6-source-code-search-engines-you-can-use-for-programming-projects/>, 2013. Acessado em Outubro/2013.
- [3] D. Bielik and S. Sonnes. Nerdyydata.
<http://nerdyydata.com/>, 2013. Acessado em Dezembro/2013.
- [4] R. G. G. Cattell and D. K. Barry. *The Object Data Standard: ODMG 3.0*. Morgan Kaufmann, 2000.
- [5] D. Chung and K. Wang. krugle.
<http://www.krugle.com/>, 2013. Acessado em Novembro/2013.
- [6] W. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Pearson Education, Boston, MA, USA, 1th edition, 2010.
- [7] C. de Faveri. Uma linguagem específica de domínio para busca em código orientado a aspectos. Mestrado, Pós Graduação em Informática, Universidade Federal de Santa Maria, 2013.
- [8] D. R. M. F. dos Santos. Recuperação de informação utilizando apache lucene e wordnet, 2012.
- [9] G. Kiczales, E. Hilsdale, J. Hugunin, M. Kersten, J. Palm, and W. G. Griswold. An overview of aspectj. In *Proceedings of the 15th European Conference on Object-Oriented Programming*, ECOOP '01, pages 327–353, London, UK, UK, 2001. Springer-Verlag.
- [10] M. McCandless, E. Hatcher, and O. Gospodnetic. *Lucene In Action*. Manning Publications Co., Stanford, CT, USA, 2th edition, 2010.
- [11] A. Sonawane. Usando o apache lucene para procura de texto.
<http://www.ibm.com/developerworks/br/java/library/os-apache-lucenesearch/>, 2009. Acessado em Maio/2013.
- [12] D. V. Winck and V. Goetten. Aspectj em 20 minutos. page 01, 2006.

Auto-parametrização do GRASP com *Path-Relinking* no agrupamento de dados com *F-Race* e *iterated F-Race*

Alternative Title: Automatic Tuning of GRASP with Path-Relinking in data clustering with F-Race and iterated F-Race

Júlio Cesar da Silva
Universidade Federal de Itajubá
Instituto de Matemática e
Computação
Av. BPS, 1303 - Pinheirinho,
Itajubá/MG. CEP 37500-903.
+55 35 9812-2582
julio_sxs@hotmail.com

Rafael de M. D. Frinhani
Universidade Federal de Itajubá
Instituto de Matemática e
Computação
Av. BPS, 1303 - Pinheirinho,
Itajubá/MG. CEP 37500-903.
+55 35 3629-1830
frinhani@unifei.edu.br

Ricardo Martins A. Silva
Universidade Federal de
Pernambuco
Centro de Informática
Cidade Universitária, Recife/PE.
CEP 50740-560.
+55 81 2126-8430
rmas@cin.ufpe.br

Geraldo Robson Mateus
Universidade Federal de Minas Gerais
Departamento de Ciência da
Computação
Av. Antônio Carlos, 6627 - ICEx
Pampulha, Belo Horizonte/MG. CEP
31270-901. +55 31 3409-5879
mateus@dcc.ufmg.br

RESUMO

Em estudos que utilizam metaheurísticas embora os parâmetros de entrada influenciem diretamente o desempenho do algoritmo sua definição na maioria das vezes é feita manualmente levantando questões sobre a qualidade dos resultados. Este artigo se propõe a aplicar o I/F-Race na auto-parametrização do GRASP com *Path-Relinking* no agrupamento de dados visando obter melhores resultados em relação aos parametrizados manualmente. Experimentos realizados com cinco conjuntos de dados demonstraram que a utilização do I/F-Race contribui para obtenção de melhores resultados que a parametrização manual.

Palavras-Chave

Auto-parametrização, I/F-Race, GRASP, *Path-Relinking*, Agrupamento de Dados.

ABSTRACT

In studies that use metaheuristics although the input parameters directly influence the performance of the algorithm its definition is mostly done manually raising questions about the quality of the results. This paper aims to apply the F/I-Race in the self parameterization of GRASP with Path-Relinking in the data clustering in order to obtain better results than the manually tuned algorithms. Experiments performed with five data sets showed that the use of I/F-race contributed to achievement best results than manual tuning.

Categories and Subject Descriptors

I.2.8 [Artificial Intelligence]: Problem Solving, Control Methods, and Search - Heuristic methods.

General Terms

Performance, Design, Experimentation.

Keywords

Automatic Tuning, I/F-Race, GRASP, *Path-Relinking*, Data Clustering.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.

Copyright SBC 2015.

1. INTRODUÇÃO

A análise de agrupamentos (*clustering*) é uma tarefa não supervisionada cujo propósito é a separação de objetos com características similares em grupos a partir de uma métrica. A quantidade de objetos e de grupos necessários para uma caracterização coerente do conjunto de dados tornam a tarefa de agrupamento um complicado problema combinatório. O ideal é que as soluções obtidas por um algoritmo de agrupamento apresentem alta homogeneidade entre os objetos de um mesmo grupo e alta heterogeneidade entre os grupos. Na literatura é possível encontrar aplicações do agrupamento de dados na mineração de dados textuais [30], mensuração e detecção de alterações da expansão urbana de regiões [36], séries temporais de dados médicos [42], agrupamento de sistemas orientados a objetos [44], entre outras.

Existe uma grande variedade de algoritmos de agrupamento cada um possuindo vantagens e desvantagens frente a tipos específicos de dados. Em diversos trabalhos as metaheurísticas têm sido utilizadas de forma isolada ou complementar aos algoritmos clássicos de agrupamento (ex. *k-means* [23], *k-medians* [26] e PAM (*Partitioning Around Medoids*) [27]), com objetivo de melhorar o desempenho e/ou precisão dos agrupamentos. Exemplos de metaheurísticas aplicadas no agrupamento de dados inclui *Simulated Annealing* [7], Algoritmos Genéticos (AG) [22], Redes Neurais Artificiais (RNAs) [12], *Fuzzy Clustering Neural Network Architecture* [37], GRASP em bases biológicas [35].

Em um dos trabalhos mais recentes a abordagem híbrida GRASP com *Path-Relinking* (GRASP+PR) proposta por [15] apresentou melhores resultados que o GRASP na sua versão tradicional. Entretanto, como em outras metaheurísticas, a definição dos parâmetros de execução do GRASP+PR é uma tarefa complexa e na maioria das vezes sua escolha é feita empiricamente. Essa abordagem geralmente resulta em uma configuração não otimizada, com possibilidades de não se explorar todas as potencialidades do algoritmo e abre margens para se questionar se os parâmetros escolhidos são os que realmente trarão melhores resultados. Nesse contexto a auto-parametrização (*automatic tuning*) de metaheurísticas, como o GRASP+PR, pode trazer vantagens como redução do tempo computacional e/ou permitir ao algoritmo alcançar melhores soluções que as versões parametrizadas manualmente.

O presente trabalho tem como objetivo aplicar a abordagem *racing* através dos métodos *F-Race* e *Iterated F-Race* (*Irace*) na auto-parametrização do GRASP+PR no problema de agrupamento

de dados. Partimos da hipótese que sua aplicação pode contribuir para melhoria do desempenho do algoritmo, redução do valor da função objetivo e/ou obtenção de agrupamentos mais coesos quando comparado a parametrização manual.

Este trabalho está organizado como se segue: A seção 2 apresenta a modelagem utilizada. A seção 3 descreve as metaheurísticas GRASP e GRASP+PR e sua aplicação no agrupamento de dados. A seção 4 descreve técnicas de auto-parametrização de metaheurísticas com maior destaque nos métodos *racing* (*F-Race* e *Irace*) que são foco de estudo desse trabalho. A seção 5 contém a metodologia utilizada e por fim a seção 6 contém os resultados dos experimentos e discussões.

2. MODELAGEM

O modelo utilizado neste trabalho segue o proposto por [35] que considera o conjunto de dados como um único grafo totalmente conexo. A distância entre cada par de objetos, obtida a partir dos atributos de cada objeto, é representada por uma aresta valorada.

Durante a execução do procedimento as arestas vão sendo gradualmente eliminadas de modo a formar cliques que representam os agrupamentos. O objetivo desta estratégia é minimizar a soma total das distâncias entre os objetos do mesmo grupo (intra-grupo) considerando que quanto menor for a distância entre dois objetos, maior a similaridade entre eles. O modelo em questão é formalizado como:

$$\sum_{i=1}^{N-1} \sum_{j=i+1}^N d_{ij} y_{ij}$$

Sujeito a:

$$\sum_{k=1}^M x_{ik} = 1, \quad i = 1, \dots, N \quad (1)$$

$$\sum_{i=1}^M x_{ik} \geq 1, \quad k = 1, \dots, M \quad (2)$$

$$x_{ik} \in \{0, 1\}, \quad i = 1, \dots, N, \quad k = 1, \dots, M \quad (3)$$

$$y_{ij} \geq x_{ik} + x_{jk} - 1, \quad i = 1, \dots, N, \quad j = i + 1, \dots, N, \quad k = 1, \dots, M \quad (4)$$

$$y_{ij} \geq 0 \quad i = 1, \dots, N, \quad j = i + 1, \dots, N \quad (5)$$

Onde d_{ij} é a distância entre os objetos i e j ; N é a quantidade de objetos do conjunto de dados, M é a quantidade de agrupamentos, x_{ik} é a variável binária que assumirá o valor 1 se o objeto pertencer ao grupo k ou 0 em caso contrário; y_{ij} é uma variável real que assume o valor 1 se os objetos i e j pertencem ao mesmo grupo. A restrição (1) garante que o objeto i pertença a apenas um grupo; (2) garante que o grupo k possua pelo menos um objeto; (3) garante que a variável x_{ik} seja binária; (4) e (5) garante que y_{ij} assume o valor 1 se ambos os valores x_{ik} e x_{jk} são 1.

3. GRASP E GRASP COM PATH-RELINKING NO AGRUPAMENTO

GRASP [13] é uma metaheurística que combina um método construtivo com busca local em um processo iterativo e independente utilizando o conceito multi-inícios. Em sua primeira fase, chamada fase construtiva, o GRASP utiliza as abordagens gulosa e aleatória para gerar soluções iniciais a partir de um processo semiguloso, e em uma segunda fase, chamada fase exploratória ou busca local, aplica em cada solução previamente construída uma busca local com o objetivo de melhorar os

resultados. GRASP foi aplicado com sucesso na resolução de diversos problemas de otimização combinatória [29, 41, 2]. [35] utilizaram o GRASP no agrupamento de dados e encontraram melhores resultados em comparação a métodos clássicos (*k-means*, *k-medians* e *PAM*).

No GRASP as soluções encontradas em iterações anteriores não influenciam o algoritmo na solução atual, ou seja, o funcionamento do GRASP tradicional não se baseia em aprendizagem sobre a sua execução. Porém é possível melhorar o GRASP através da utilização de mecanismos de memória que permitam um direcionamento do algoritmo para as soluções mais promissoras encontradas previamente.

Path-Relinking [17, 18, 19] é um procedimento de intensificação de busca que tem sido aplicado com sucesso na incorporação de memória ao GRASP. O método trabalha explorando as trajetórias que conectam soluções de alta qualidade (soluções elite), obtidas durante as execuções do GRASP. O procedimento parte de uma solução inicial e gera um caminho de soluções no espaço de vizinhança que conduz a uma solução guia. O objetivo é encontrar soluções diversificadas e de qualidade entre duas soluções elite.

Existem variações do *Path-Relinking* que possuem estratégias diferentes na determinação da solução inicial/guia, bem como a trajetória a ser seguida [39, 40, 2]. Na variante *Forward* a pior solução em termos de função objetivo é tomada como solução inicial e a melhor como guia. Na *Backward* considera-se a pior solução como guia e a melhor como solução inicial. Na *Mixed* o início da trajetória é alternada entre a melhor solução e a pior até que seja encontrada uma solução intermediária. Com a *Greedy Randomized Adaptive* uma solução guia é selecionada aleatoriamente de uma lista de candidatos mais promissores. Por fim, a *Truncated* explora uma parte da trajetória entre as soluções.

Algoritmo 1: GRASP+PR()

```

Data:  $I, M, TP, p, DS, Ism, DC, S$ ;
Result: Solução  $x^* \in X$ ;
1.  $P := \emptyset$ ;
2. while critério de parada não alcançado do
3.    $x' := \text{FaseConstrutiva}()$ ;
4.   if  $P$  possuir ao menos  $p$  elementos then
5.      $x' := \text{FaseBuscaLocal}(x')$ ;
6.     Aleatoriamente seleciona solução  $y \in P$ ;
7.      $x' := \text{PathRelinking}(x', y)$ ;
8.      $x' := \text{FaseBuscaLocal}(x')$ ;
9.     if  $P$  estiver completo then
10.      if  $c(x') \leq \max\{c(x) \mid x \in P\}$  and  $x' \neq P$  then
11.        Substitui elemento mais similar a  $x'$ 
        entre todos com custo pior que  $x'$ ;
12.      end
13.     else if  $x' \neq P$  then
14.        $P := P \cup \{x'\}$ ;
15.     end
16.   else if  $x' \neq P$  and  $P = \emptyset$  then
17.      $P := P \cup \{x'\}$ ;
18.   end
19. end
20. end
21. return  $x^*$ 

```

O Algoritmo 1 [15] descreve o GRASP+PR no agrupamento de dados, que toma como entrada o conjunto de dados I e os parâmetros: M que é a quantidade agrupamentos, TP é o tamanho do conjunto de soluções elite P , p é a quantidade mínima de elementos em P antes do início do *PR*, DS é a diferença simétrica entre a solução inicial e a guia, Ism é a quantidade de iterações sem melhoria como critério de parada do algoritmo, DC é a

medida de similaridade entre os objetos do conjunto de dados e S a semente utilizada pelo método aleatório.

Após inicializar o conjunto elite P como vazio (linha 1), o GRASP+PR é executado nas linhas 2 a 19 até que um critério de parada (Ism) seja satisfeito. Em cada iteração uma solução x' é gerada pela fase construtiva na linha 3. Na linha 4 é verificada a quantidade de soluções pertencentes a P . Caso P esteja vazio, a solução x' é inserida no conjunto. Se P não possuir no mínimo p elementos e se a solução x' for suficientemente diferente de todas as soluções presentes em P , então x' é inserido em P (linha 16). Se P possuir pelo menos p elementos, então os passos 5 à 15 são executados. A linha 5 realiza a busca local objetivando a melhoria da solução x' construída na Fase Construtiva (linha 3). Uma solução $y \in P$ é escolhida aleatoriamente para aplicação do PR com a solução x' (linhas 6 e 7).

Após a realização do *Path-Relinking*, a solução x' passa novamente pela Fase Busca Local visando melhorias (linha 8). Caso o conjunto elite esteja completo é verificado se a solução é suficientemente diferente das soluções atualmente presentes em P ($x' \neq P$). Esta estratégia visa manter em P soluções de qualidade e diversificadas. Ao se respeitar este critério, a solução x' se torna candidata a entrar no conjunto elite substituído desse conjunto a solução x mais similar a x' , isto é, a que apresente a menor diferença simétrica $\Delta(x, x')$. Se P não estiver completo a solução x' é inserida caso atenda a restrição $x' \neq P$. Ao final, o algoritmo retorna x^* que armazena a melhor solução encontrada. Detalhes da *FaseConstrutiva()*, *FaseBuscaLocal()* e *PathRelinking()* podem ser obtidos em [15].

4. AUTO-PARAMETRIZAÇÃO DE METAHEURÍSTICAS

A parametrização de metaheurísticas, como GRASP+PR, envolve a definição de valores que serão utilizados como parâmetros de entrada para execução do algoritmo. Geralmente existe uma grande variação desses valores caracterizando esta tarefa como um típico problema combinatório. Na maioria dos trabalhos envolvendo metaheurísticas a definição dos parâmetros é realizada de forma manual e empírica levantando questões sobre o desempenho do algoritmo a partir dos parâmetros escolhidos, ou seja, se realmente são os que trarão melhores resultados. A auto-parametrização de metaheurísticas (*Automatic Tuning*) tem recebido atenção por possibilitar a obtenção de melhores soluções em comparação a parametrização manual.

[11] apresentou um *framework* para auto-parametrização de metaheurísticas combinando projetos de experimentos e Redes Neurais Artificiais. [14] utilizaram a variante do algoritmo genético, *Biased Random Key Genetic Algorithm* (BRKGA) [21], para auto-parametrização do GRASP com *Path-Relinking* no *Generalized Quadratic Assignment Problem* (GQAP). [6] realizou a auto-parametrização das metaheurísticas Colônia de formigas e Busca Local iterativa utilizando *F-Race*. [31] implementaram o *iterated F-Race* (*Irace*) visando corrigir falhas do *F-Race*. *IF-Race* utilizam uma abordagem estatística para selecionar a melhor configuração num conjunto de configurações candidatas [4].

4.1 A abordagem *racing*

De acordo com [5] a abordagem *racing* teve início quando [32] tiveram problemas com a seleção de modelos para aprendizado de máquina. [4] adaptaram essa abordagem para empregá-la em trabalhos envolvendo a configuração de algoritmos. Essa abordagem consiste na execução de várias etapas para avaliação de um conjunto finito de configurações candidatas, onde após

testes estatísticos em cada uma destas etapas, as configurações menos eficientes são descartadas permitindo desta maneira concentrar os cálculos nas configurações candidatas mais promissoras.

A abordagem *racing* considera uma sequência de instâncias i_k , com $k=1, 2, 3, \dots$, onde cada instância representa um conjunto de parâmetros gerados de forma aleatória. Tem-se como c_k^θ o custo de um conjunto de parâmetros θ em uma instância i_k . A avaliação dos conjuntos de parâmetros candidatos é realizada de forma incremental de modo que no passo k obtém-se um vetor de custos (6) que serão utilizados para avaliar θ .

$$c_k(\theta) = (c_1^\theta, c_2^\theta, \dots, c_k^\theta) \quad (6)$$

Uma matriz de custos é obtida unindo-se c_k^θ aos vetores de custos finais de uma etapa anterior, dados por $c^{k-1}(\theta)$. Em seguida é gerada a próxima sequência de conjunto de parâmetros candidatos, onde θ_k é o conjunto dos conjuntos de parâmetros sobreviventes antes do passo k (7).

$$\theta_0 \supseteq \theta_1 \supseteq \theta_2 \supseteq \dots, \quad (7)$$

4.1.1 F-RACE

O *F-Race* [4] utiliza uma abordagem *racing* baseada no teste de Friedman [8] que analisa e classifica os custos levando em consideração a variância encontrada. De acordo com [5] o funcionamento do *F-Race* supõe que estão no processo $m=|\theta_{k-1}|$ conjuntos de parâmetros candidatos. O teste de Friedman assume que os k custos observados, são agrupados em b_i blocos independentes (8), com $i = 1, 2, \dots$

$$\begin{aligned} b_1 &= (c_1^{\theta_{v1}}, c_1^{\theta_{v2}}, \dots, c_1^{\theta_m}) \\ b_2 &= (c_2^{\theta_{v1}}, c_2^{\theta_{v2}}, \dots, c_2^{\theta_m}) \\ &\vdots \\ b_k &= (c_k^{\theta_{v1}}, c_k^{\theta_{v2}}, \dots, c_k^{\theta_m}) \end{aligned} \quad (8)$$

Nos blocos, os custos c_l^θ são organizados em ordem crescente e para cada conjunto de parâmetros $\theta_{v_j} \in \theta_{k-1}$, R_{lj} é a classificação de θ_{v_j} no bloco b_l , e $R_j = \sum_{l=1}^k R_{lj}$ é a soma da classificação de θ_{v_j} , sobre todas as instâncias i_l , com $1 \leq l \leq k$. O teste de Friedman utilizado está descrito em (9) [8]:

$$T = \frac{(m-1) \sum_{j=1}^m (R_j - \frac{k(m+1)}{2})^2}{\sum_{l=1}^k \sum_{j=1}^m R_{lj}^2 - \frac{km(m+1)^2}{4}} \quad (9)$$

Sob a hipótese nula de que todos os candidatos são equivalentes, T é aproximadamente χ^2 de uma distribuição com $m-1$ graus de liberdade [38]. Se o valor observado de T é maior do que $1-\alpha$ desta distribuição, onde α é o nível de confiança escolhido, a hipótese nula é rejeitada o que indica que ao menos um dos conjuntos de parâmetros candidatos apresenta melhor desempenho do que as demais. Após o teste de hipótese, o teste *post hoc* de Friedman (10) é utilizado visando encontrar o conjunto de parâmetros mais promissor e quais deverão ser eliminados [8].

Se o *F-Race* não rejeita a hipótese nula da comparação, então todas as configurações candidatas de θ_{k-1} passam para θ_k . Os demais conjuntos de parâmetros que não possuem um bom nível de significância são descartados e não mais aparecem em θ_k .

$$\sqrt{\frac{|R_j - R_{jh}|}{2k(1 - \frac{T}{k(m-1)})(\sum_{i=1}^k \sum_{j=1}^m R_{ij}^2 - \frac{km(m+1)^2}{4})}} > t_{1-\alpha/2}, \quad (10)$$

O Algoritmo 2 descreve o método *F-Race* [3] que toma como entrada o conjunto de configurações *C* que serão executadas no conjunto de dados *I* e a quantidade mínima de custos a serem analisados. Tem como resultado o melhor conjunto de parâmetros. Na linha (1) são inicializados os atributos que fornecem a melhor configuração e o número mínimo de custos. Na estrutura de repetição da linha (2) são executados todos os conjunto de parâmetros no conjunto de dados. Já na linha (6) é utilizado o teste de Friedman com base na classificação dos custos para um número de conjuntos de parâmetros independentes ocorridos até o momento e um número de conjuntos de parâmetros ainda no processo. Em seguida na linha (10) são realizadas comparações em pares com os melhores conjuntos de parâmetros até o momento. E por fim na linha (12) são eliminados os conjuntos de parâmetros não promissores. Este procedimento deve ser executado até que exista apenas um conjunto de parâmetros ou quando se atinja uma quantidade máxima de iterações definidas pelo usuário.

Algoritmo 2: *F-Race*()

Data: Conjunto de Configurações *C*;
Conjunto de Dados *I*;
Número de Conjuntos Min *ni_{min}*;
Result: Melhor configuração *C**;
1. *C**:=*C*; *ni*:=0;
2. **repeat**
3. rodar as configurações de *C** em *i*;
4. *ni*:=*ni*+1;
5. **if** *ni* ≥ *ni_{min}* **then**
6. Executar o Teste de Friedman
7. **if** for indicado uma diferença de Desempenho significativa **then**
8. *c**:= retornar melhor Configuração em *C**
9. **for** *c* ∈ *C** \ {*c**} **do**
10. Executar o teste pairwise Friedman post hoc em *c* e *c**;
11. **if** for encontrado um diferença de desempenho significativo **then**
12. eliminar *c* de *C**;
13. **end if**;
14. **end for**;
15. **end if**;
16. **end if**;
17. **until** a condição seja satisfeita;
18. **return** *C**
19. **end**

4.1.2 IRACE

[24] mostra que uma grande limitação do *F-Race* é que nos passos iniciais todas as configurações precisam ser avaliadas. Esta característica do *F-Race* limita consideravelmente o tamanho do conjunto de parâmetros para que o procedimento possa ser aplicado de forma eficaz, especialmente quando se trata de um conjunto de parâmetros que contém diversas combinações de valores.

Iterated F-Race (Irace) [5] é uma variante do *F-Race* que tem o objetivo de resolver os problemas da sua limitação. *Irace* utiliza o procedimento do *F-Race* descrito na seção 4.1.1 como uma sub-rotina, onde em cada iteração uma quantidade de conjuntos de parâmetros sobrevive servindo para gerar uma amostragem de

novos conjuntos de parâmetros. Espera-se deste modo, gerar amostras dos conjuntos de parâmetros mais promissores. *Irace* segue diretamente uma estrutura de pesquisa baseado em modelo que normalmente é aplicado em três etapas [43]: (1) Geração de amostras de conjuntos de parâmetros de acordo com uma distribuição específica, (2) Seleção dos melhores conjuntos obtidos por meio do algoritmo *Irace*, e (3) Atualização das amostras removendo conjuntos de parâmetros de pior desempenho de modo a enviesá-las. Estes três passos são repetidos até que um critério predefinido seja satisfeito.

O Algoritmo 3 descreve o método *Irace*. Na linha (1) tem-se o conjunto de configurações que receberá os melhores conjuntos de parâmetros candidatos encontradas pelo *F-Race*. Na linha (3) são gerados novos conjuntos baseados nas melhores configurações encontradas pelo processo do *F-Race* até o momento. Já na linha (7) é retornada o melhor conjuntos de parâmetros do conjunto elite. Este procedimento continuará até atingir um número mínimo de conjuntos de parâmetros sobreviventes ou um número máximo de iterações.

Algoritmo 3: *Irace*()

Data: Conjunto de Configurações *C*;
Result: Conjunto Elite *C**;
1. *C'*:=0; // *C'* é inicializado a partir de uma //execução do *F-Race*
2. **repeat**
3. Realiza uma amostragem a partir de *C** para gerar novas configurações;
4. Executa as configurações no *F-Race*;
5. Adiciona as melhores configurações em *C**;
6. **until** a condição seja satisfeita;
7. **return** *C** ∈ *C** //É a melhor configuração
8. **end**

5. METODOLOGIA

Este trabalho se propõe a estender os estudos realizados por [15] e através da aplicação do *I/F-Race* realizar a auto-parametrização do GRASP+PR aplicado ao agrupamento de dados. Parte-se da hipótese que a utilização do *I/F-Race* na auto-parametrização do GRASP+PR no problema de agrupamento de dados pode contribuir para melhoria do desempenho, diminuição da função objetivo e/ou obtenção de agrupamentos mais coesos quando comparados com a parametrização manual. A Figura 1 ilustra a sequência de etapas que envolvem a execução do *F-Race* na auto-parametrização do GRASP+PR.

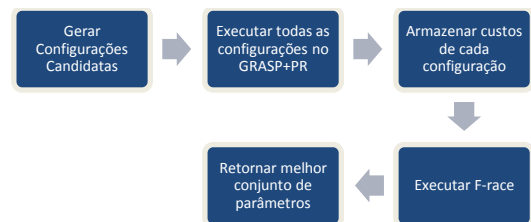


Figura 1 – Etapas do *F-Race* na auto-parametrização do GRASP+PR.

Na primeira etapa são gerados os conjuntos de parâmetros, seguida da execução destes parâmetros no GRASP+PR e do armazenamento dos custos obtidos a partir de cada conjunto de parâmetros. Por fim, após a execução do *F-Race*, o algoritmo retorna o conjunto de parâmetros que forneceu o melhor custo.

A Figura 2 ilustra a sequência de etapas para a execução do *Irace*. A primeira parte é essencialmente o *F-Race*, a diferença é que o melhor conjunto de parâmetros é adicionado a um conjunto elite e

novos conjuntos de parâmetros são gerados a partir desse conjunto. Este ciclo é realizado até que uma condição de parada seja satisfeita e o último conjunto elite é retornado.

As implementações deste trabalho, consideraram a criação de um vetor que contém todos os custos ordenados de forma decrescente, obtidos a partir da execução do GRASP+PR com cada conjunto de parâmetros. O GRASP+PR foi executado três vezes para cada conjunto de parâmetros onde se obteve a média dos valores de CRand, Função objetivo e Tempo de execução. Essa quantidade foi considerada visando reduzir o tempo de execução do GRASP+PR e se mostrou suficiente para obtenção dos melhores resultados. O critério de melhoria foi baseado na seguinte prioridade: (1) maior valor médio do CRand, (2) menor valor médio da função objetivo e (3) menor tempo de execução.

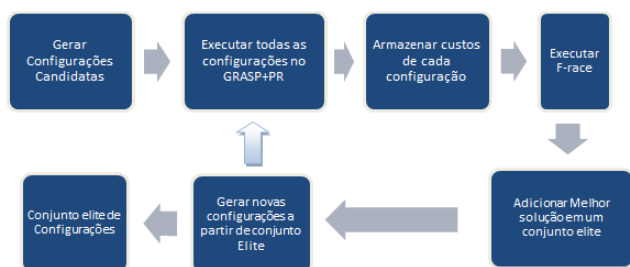


Figura 2 – Aplicação do *I/race* na auto-parametrização do GRASP+PR.

Para executar o teste de Friedman, o vetor de custos precisa ser dividido em uma matriz com no mínimo 2 linhas para que seja montada uma tabela onde será feita a etapa de criação do *ranking* conforme proposto por [8]. A partir dessa definição foi utilizada a seguinte estratégia, a quantidade de custos do vetor obtidos na primeira etapa foi dividida em uma matriz com a quantidade de 15 colunas, denominadas grupos e N/15 linhas denominadas blocos, sendo N a quantidade de conjuntos de parâmetros. Para a realização dos experimentos os métodos *I/F-race* foram executados considerando 60, 90, 150 e 210 conjuntos de parâmetros. Estes valores seguem as definições de populações iniciais nos métodos utilizados em [28] e tem como objetivo verificar se a quantidade de conjunto de parâmetros influencia na obtenção de melhores configurações.

A partir da matriz de custos é construída uma nova matriz contendo o *ranking* de todos valores contidos nos blocos. Nesta etapa é realizado o test *post hoc* de Friedman (10) com nível de confiança de $\alpha=0,05$ para realizar o cálculo das diferenças observadas. Em seguida, com base na tabela da distribuição T, é verificado quais grupos possuem diferença maior que a diferença crítica, o que fornece subsídios para eliminação dos piores grupos.

Os conjuntos de parâmetros que restaram após as eliminações são reordenados para permitir que os blocos que perderam grupos possam ser completados com os conjuntos de parâmetros sobreviventes de outros blocos. Este balanceamento faz com que o número de blocos diminua gradualmente com as execuções do algoritmo. Os passos anteriores são repetidamente realizados até que não exista mais diferenças entre os grupos e ao final a primeira posição do pool de conjunto parâmetros, que possui o melhor custo, é retornada.

4.1. Métricas Analisadas

Os algoritmos foram analisados considerando o tempo de execução (em segundos.millessegundos), *Time to Target Plot*

(*ttplot*) [1], função objetivo (FO) e *Corrected Rand Index* (CRand) [25]. O *ttplot* é uma medida utilizada para comparação de tempo em algoritmos estocásticos. Para construção dos *ttplots* os algoritmos foram executados 200 vezes com os parâmetros manuais e os encontrados pelos métodos *I/F-Race*. A função objetivo do GRASP+PR fornece uma referência para coesão dos agrupamentos gerados e por fim o CRand compara a similaridade entre os agrupamentos obtidos com os agrupamentos de referência fornecidos com o conjunto de dados. A validação dos experimentos se deu pela comparação dos resultados obtidos com *I/F-Race* na auto-parametrização do GRASP+PR nas variantes *forward*, *backward*, *mixed*, *greedy randomized adaptive* e *truncated*, com as mesmas variantes parametrizadas manualmente.

4.2. Ambiente de desenvolvimento e testes

O *I/F-Race* foi implementado em ambiente *Microsoft Windows* utilizando a interface de desenvolvimento *NetBeans 7.1.2* e linguagem *Java SE Ver. 7u21*. Foram utilizadas as seguintes bibliotecas: *Colt* ver. 1.0.3 [33] para geração de números aleatórios, *CILIB* [10], *Functional Java* [34], *Google-collect-1.0* [20] para realização dos cálculos referentes ao teste de Friedman, e *Renjin* [16] para executar funções da linguagem R no Java como a função "tnorm" [9] que gera uma amostragem de acordo com distribuição normal truncada. Os experimentos foram realizados em um *notebook HP pavillion g4-2220br*, com um processador *Core i3-3110M CPU @ 2.40 GHz* e 6 GB de memória RAM, executando o sistema operacional *Microsoft Windows 8* e a máquina virtual *Java SE Ver. 7u45*.

Tabela 1 – Descrição dos parâmetros e respectivos limites.

Parâmetro	Descrição	Min	Máx
TP	Tamanho do <i>pool</i> . Número de soluções que o <i>Path-Relinking</i> irá armazenar no conjunto elite.	2	30
IPR	Número mínimo de soluções presentes no <i>pool</i> antes do início do <i>Path-Relinking</i> .	2	TP
DS	Diferença simétrica entre uma nova solução e uma solução presente no <i>pool</i> como critérios para que a nova solução possa ser adicionada ao <i>pool</i> de soluções.	2	50% do conjunto de dados.
ISM	Quantidade máxima de iterações sem melhorias do GRASP+PR para que este encerre seu funcionamento.	1	150
BL	Máximo de iterações na fase de busca local do GRASP.	1	300
NH	Porcentagem do tamanho da lista preliminar de arestas nos conjuntos de dados que contêm mais de 15 objetos.	10	50
QTDGRUPOS	Quantidade de grupos utilizados na tarefa de agrupamento.	2	30
VAR_PR	Variante <i>Path-Relinking</i> usada (0= <i>Forward</i> , 1= <i>Backward</i> , 2= <i>Mixed</i> , 3= <i>Greedy Randomized</i> , 4= <i>Truncated</i>).	0	4

O código do GRASP+PR utilizado segue o Algoritmo 1 [15]. Os parâmetros que serão auto-parametrizados, foram escolhidos com base na sua influência no comportamento do algoritmo. Os parâmetros considerados nos experimentos estão descritos na Tabela 1.

A distância *City Block* (11) foi utilizada para o cálculo da similaridade entre os objetos, onde, d_{ij} é a similaridade entre os objetos i e j , L é a quantidade de objetos do conjunto de dados, a_{ik} e a_{jk} são respectivamente o k -ésimo atributo dos i -ésimo e j -ésimo objetos a_i e a_j .

$$d_{ij} = \sum_{k=1}^L |a_{ik} - a_{jk}| \quad (11)$$

A Tabela 2 contém detalhes dos conjuntos de dados utilizados. Os resultados com parâmetros manuais foram obtidos nos estudos realizados por [15].

Tabela 2 – Conjuntos de dados utilizados nos experimentos.

Conjunto de Dados	#Objetos	#Atributos	Referência
BreastB2	49	1213	[45]
DLBCLA	141	661	[45]
MultiA	103	5565	[46]
Novartis	103	1000	[46]
Glass	214	9	[47]

6. RESULTADOS E DISCUSSÃO

Os resultados obtidos com os experimentos estão descritos na Tabela 3. Os parênteses nas linhas *I/F-Race* contém a variante do PR e a quantidade de conjuntos de parâmetros que apresentaram os melhores resultados, os quais estão destacados em negrito.

É possível observar a ocorrência de maiores CRands em três dos cinco conjunto de dados analisados, tendo o *F-Race* contribuído para melhorias nos conjunto de dados BreastB2 e DLBCLA, e o *Irace* para o MultiA. A redução do desvio padrão no BreastB2, DLBCLA e Glass, mostram uma melhoria na robustez do método.

Com relação a função objetivo, cuja redução indica uma maior coesão dos agrupamentos, embora tenha-se observado melhoria apenas no conjunto de dados BreastB2, é possível constatar reduções significativas no desvio padrão em quatro dos cinco conjunto de dados BreastB2, MultiA, DLBCLA e Glass.

Melhorias no tempo de execução são observadas nos conjunto de dados MultiA, Novartis e DLBCLA. No Novartis, o *Irace* apresentou o melhor tempo sem comprometer o valor do CRand. Como valor *target* para o *ttplot*, considerou-se o maior valor da função objetivo obtido em cada conjunto de dados.

Os gráficos *ttplot* das figuras 4 e 5 ilustram melhorias de tempo no Novartis e DLBCLA respectivamente, onde é possível constatar que nos algoritmos auto-parametrizados o tempo de convergência para alcançar o valor da função objetivo é significativamente menor que os parametrizados manualmente.

A análise dos dados confirma a hipótese que a auto-parametrização do GRASP+PR no agrupamento de dados através do *I/F-Race* contribui para melhoria dos resultados quando comparados aos algoritmos parametrizados manualmente.

Tabela 3 – Resumo dos resultados obtidos nos experimentos.

ALGORITMO	CRAND		FO		TEMPO (s.ms)	
	Média	DP	Média	DP	Média	DP
Breast B2						
<i>F-Race</i> (PRb/210)	0.322	0.0372	68261.9	58.927	4.342	4.630
<i>Irace</i> (PRb/90)	0.311	0.0389	68403.1	154.063	1.099	1.223
GRASP-PRb	0.285	0.0478	68798.2	334.293	0.424	0.458
GRASP-PRf	0.266	0.0552	69232.6	431.619	0.379	0.408
GRASP-PRgr	0.236	0.0612	69917.8	507.777	0.197	0.220
GRASP-PRm	0.243	0.0617	69831.4	476.933	0.276	0.299
GRASP-PRt	0.266	0.0552	69232.6	431.619	0.345	0.381
MultiA						
<i>F-Race</i> (PRgr/150)	0.869	0.0227	1.48E+09	222774.000	6.316	6.568
<i>Irace</i> (PRm/60)	0.880	0.0415	1.48E+09	847452.000	0.084	0.096
GRASP-PRb	0.861	0.0178	1.48E+09	290116.000	1.527	1.594
GRASP-PRf	0.864	0.0168	1.48E+09	339546.000	1.386	1.477
GRASP-PRgr	0.878	0.0157	1.48E+09	297480.000	1.564	1.653
GRASP-PRm	0.879	0.0169	1.48E+09	322229.000	1.369	1.429
GRASP-PRt	0.864	0.0168	1.48E+09	339546.000	1.399	1.468
Novartis						
<i>F-Race</i> (PRm/210)	0.944	0.026	1033170	344.674	0.114	0.125
<i>Irace</i> (PRf/210)	0.950	0.0000	1033080	0.000	0.489	0.517
GRASP-PRb	0.950	0.0000	1033080	0.000	1.889	2.043
GRASP-PRf	0.950	0.0000	1033080	0.000	2.195	2.364
GRASP-PRgr	0.950	0.0000	1033080	0.000	2.095	2.282
GRASP-PRm	0.950	0.0000	1033080	0.000	2.431	2.607
GRASP-PRt	0.950	0.0000	1033080	0.000	2.231	2.400
DLBCLA						
<i>F-Race</i> (PRm/60)	0.834	0.0111	4.61E+08	5413.550	2.537	2.630
<i>Irace</i> (PRm/210)	0.826	0.0192	4.61E+08	11669.400	0.968	1.069
GRASP-PRb	0.829	0.0174	4.61E+08	10493.400	2.136	2.339
GRASP-PRf	0.828	0.0183	4.61E+08	11398.800	2.318	2.557
GRASP-PRgr	0.822	0.0212	4.61E+08	13845.000	1.989	2.226
GRASP-PRm	0.822	0.0220	4.61E+08	14841.300	1.874	2.074
GRASP-PRt	0.828	0.0183	4.61E+08	11398.800	2.478	2.711
Glass						
<i>F-Race</i> (PRb/60)	0.260	0.0000	10173.3	3.28E-11	19.671	20.724
<i>Irace</i> (PRb/90)	0.260	0.0000	10173.3	3.28E-11	17.028	17.892
GRASP-PRb	0.260	0.0014	10173.3	2.81E-02	8.829	9.463
GRASP-PRf	0.260	0.0010	10173.3	2.00E-02	11.743	12.423
GRASP-PRgr	0.259	0.0033	10173.3	6.53E-02	7.781	8.204
GRASP-PRm	0.259	0.0029	10173.3	5.89E-02	8.206	8.658
GRASP-PRt	0.260	0.0010	10173.3	2.00E-02	9.186	9.636

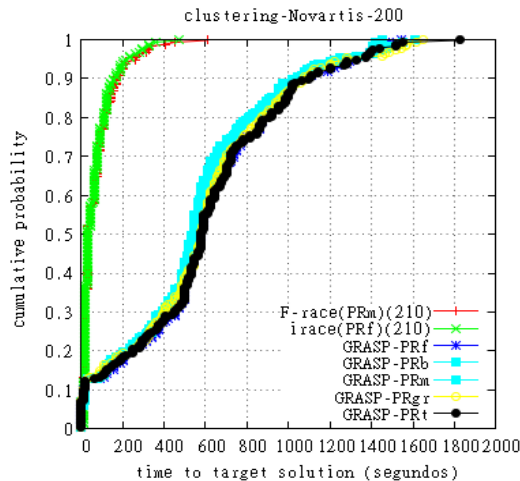


Figura 4 – tttplot do conjunto de dados Novartis

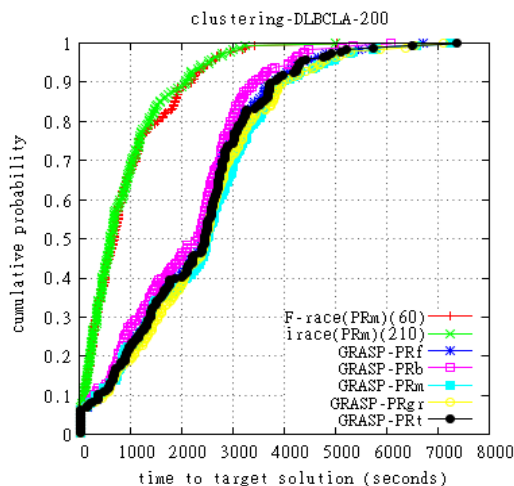


Figura 5 – tttplot do conjunto de dados DLBCLA

Pesquisas envolvendo a utilização de métodos evolutivos na auto-parametrização do GRASP+PR no agrupamento estão sendo concluídas e possibilitarão uma comparação com os métodos apresentados neste trabalho.

7. REFERÊNCIAS

- [1] Aiex, R. M.; Resende, M. G. & Ribeiro, C. C. (2007). "TTT plots: a perl program to create time-to-target plots". *Optimization Letters* 1.4, pp. 355-366.
- [2] Binato, S.; de Oliveira, G. & de Araujo, J. (2002). A greedy randomized adaptive search procedure for transmission expansion planning. *IEEE Transactions on Power Systems*, 16(2) pp. 247-253.
- [3] Birattari, M. & Dorigo, M. (2004). "The problem of tuning metaheuristics as seen from a machine learning perspective." PhD thesis, Université Libre de Bruxelles, Brussels, Belgium.
- [4] Birattari, M.; Stützle, T; Paquete, L. & Varrentrapp, K. (2002). "A Racing Algorithm for Configuring Metaheuristics." *GECCO*. Vol. 2.
- [5] Birattari, M.; Yuan, Z; Balaprakash, P. & Stützle, T. (2010). "F-Race and iterated F-Race: An overview." *Experimental methods for the analysis of optimization algorithms*, pp. 311-336, Springer Berlin Heidelberg.
- [6] Birattari, M.; Stützle, T; Paquete, L. & Varrentrapp, K. (2009) "Tuning metaheuristics: a machine learning perspective". Springer, Vol. 197.
- [7] Brown, D. E. & Huntley, C. L. (1992). A practical application of simulated annealing to clustering. In *Pattern Recognition*, pp. 401-412.
- [8] Conover, W. J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, New York: John Wiley & Sons.
- [9] Christopher, J., *Function to sample according to a truncated normal distribution*. 2011. <https://github.com/cran/Irace/blob/master/R/tnorm.R>. Acessado em 22/10/2014.
- [10] CIRG - Computational Intelligence Research Group (CIRG@UP) Department of Computer Science University of Pretoria South Africa. Biblioteca CILIB. 2010. <http://www.cilib.net/>. Acessado em 22/10/2014.
- [11] Dobslaw, F. (2010). A parameter tuning framework for metaheuristics based on design of experiments and artificial neural networks. In *Proceeding of the International Conference on Computer Mathematics and Natural Computing 2010*. WASET.
- [12] Du, K.-L. (2010). *Clustering: A neural network approach*, *Neural Networks*, Volume 23, Issue 1, pp. 89- 107.
- [13] Feo, T. A. & Resende, M. G. (1995). Greedy randomized adaptive search procedures. *Journal of global optimization*, 6(2), pp 109-133.
- [14] Festa, P.; Gonçalves, J. F.; Resende, M. G. & Silva, R. M. (2010). "Automatic tuning of GRASP with path-relinking heuristics with a biased random-key genetic algorithm". In *Experimental Algorithms*, pp. 338-349. Springer Berlin Heidelberg.
- [15] Frinhan, R. M. D.; Silva, R. M. A. & Mateus, G. R. (2011). "GRASP com Path-Relinking para agrupamento de dados biológicos". Dissertação de Mestrado. Universidade Federal de Minas Gerais, Departamento de Ciência da Computação.
- [16] FSF - Free Software Foundation. Biblioteca Renjin, 2007. <http://www.renjin.org/>. Acessado em 22/10/2014.
- [17] Glover, F. (1997). "Tabu search and adaptive memory programming: advances, applications and challenges. In *Interfaces in computer science and operations research*" pp. 1-75. Springer US.
- [18] Glover, F.; Laguna, M. & Marti, R. (2000). *Fundamentals of scatter search and Path-Relinking*. In *Control and Cybernetics*, pp. 653-684.
- [19] Glover, F. & Marti, R. (2006). "Tabu search". In *Metaheuristic procedures for training neural networks*, pp. 53-69.
- [20] Google Inc. Biblioteca Google-Collections .2007. <http://code.google.com/p/google-collections/>. Acessado em 22/10/2014.
- [21] Gonçalves, J. F. & Resende, M. G. (2011). Biased random-key genetic algorithms for combinatorial optimization. *Journal of Heuristics*, 17(5), 487-525.

- [22] Hall, L.; Ozyurt, B. & Bezdek, J. (1999). *Clustering with a genetically optimized approach*. In *Evolutionary Computation*, pp 103–112.
- [23] Hartigan, J. A. & Wong, M. A. (1979). “Algorithm AS 136: A k-means clustering algorithm.” *Applied statistics*, pp 100-108.
- [24] Hamadi, Y.; Monfroy, E. & Saubion, F. (2012) “*Autonomous Search*” XVI, 308 p., ISBN: 978-3-642-21433-2, Springer.
- [25] Hubert, L. & Arabie, P. (1985). *Comparing partitions*. *Journal of Classification*, 2(1):193-218.
- [26] Jain, A. K. & Dubes, R. C. (1988). “*Algorithms for clustering data*.”. Englewood Cliffs. Prentice-Hall.
- [27] Kaufman, L. & Rousseeuw, P. J. (1990). “Partitioning around medoids (program pam).” *Finding groups in data: an introduction to cluster analysis*, pp 68-125.
- [28] Morán-Mirabal, L. F.; González-Velarde, J. L. & Resende, M. G. (2013). *Automatic tuning of GRASP with evolutionary path-relinking*. In *Hybrid Metaheuristics* (pp. 62-77). Springer Berlin Heidelberg.
- [29] Lee, D.; Chen, J. & Cao, J. (2010). “*The continuous berth allocation problem: A greedy randomized adaptive search solution*”. *Transportation Research Part E: Logistics and Transportation Review*, 46(6):1017-1029.
- [30] Lopes, M. C. S. (2004). *Mineração de dados textuais utilizando técnicas de clustering para o idioma português*. *Doctoral dissertation*, Departamento de Engenharia Civil. Universidade Federal do Rio de Janeiro.
- [31] López-Ibáñez, M.; Dubois-Lacoste, J.; Stützle, T.; & Birattari, M. (2011). *The irace package: Iterated racing for automatic algorithm configuration*. Tech. Rep. TR/IRIDIA/2011-004, Université Libre de Bruxelles, Belgium.
- [32] Maron, O. & Moore, A.W. (1994). *Hoeffding races: Accelerating model selection search for classification and function approximation*. In: Cowan J. D., et al. (eds), *Advances in Neural Information Processing Systems* Morgan Kaufmann, San Francisco, CA, USA, vol 6, pp. 59–66.
- [33] Matsumoto, M. & Nishimura, T. (1998). *Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator*. *ACM Transactions on Modeling and Computer Simulation*, 8:3-30.
- [34] Morris, T.; Bjarnason, R.; Adams, T; Clow, B; Clarkson, R; Partridge, N. & Zaugg, T. (2010) *Biblioteca functionaljava*. Regents of the University of California. Site: <http://www.functionaljava.org/>. Acessado em 22/10/2014.
- [35] Nascimento, M.; Toledo, F. & de Carvalho, A. (2010). *Investigation of a new grasp based clustering algorithm applied to biological data*. In *Computers and Operations Research*, pp 1381–1388.
- [36] Ohata, A. T.; & Quintanilha, J. A. (2005). O uso de algoritmos de clustering na mensuração da expansão urbana e detecção de alterações na Região Metropolitana de São Paulo. *Anais do XII Simpósio Brasileiro de Sensoriamento Remoto, Goiânia, Brasil, 16–21 de Abril de 2005*, 647-655.
- [37] Ozbay, Y.; Ceylan, R. & Karlik, B. (2006). *A fuzzy clustering neural network architecture for classification of ecg arrhythmias*. In *Computers in Biology and Medicine*, pp 376–388.
- [38] Papoulis, A. (1991) & *Random Variable Probability. "Stochastic Processes."* McGraw-Hill.
- [39] Resende, M. G. & Ribeiro, C. C.(2005). *GRASP with path-relinking: Recent advances and applications*. In *Metaheuristics: progress as real problem solvers*, pp. 29-63. Springer US.
- [40] Resende, M. G. C.; Ribeiro, C. C.; Glover, F. & Martí, R. (2010b). *Scatter search and path-relinking: Fundamentals, advances and applications*. *Handbook of Metaheuristics*, pp. 87-107.
- [41] Sirdey, R.; Carlier, J. & Nace, D. (2010). *A GRASP for a resource-constrained scheduling problem*. *International Journal of Innovative Computing and Applications*, 2(3):143-149.
- [42] Spolaôr, N.; Lee, H. D.; Ferrero, C. A.; Coy, C. S. R., Fagundes, J. J., Wu, F. C.; ... & de Coloproctologia, S. (2008). Um Estudo da Aplicação de Clustering de Séries Temporais em Dados Médicos.
- [43] Zlochín, M.; Birattari, M.; Meuleau, N. & Dorigo M. (2004) “*Model-based search for combinatorial optimization: A critical survey*.” *Annals of Operations Research*, vol 131. pp 373-395.
- [44] Botelho, A. L. V.; Semaan, G. S. & Ochi, L. S. (2011) “*Agrupamento de Sistemas Orientados a Objetos com Metaheurísticas Evolutivas*” *Anais do VII Simpósio Brasileiro de Sistemas de Informação*. págs. 153 – 165.
- [45] Monti, S.; Savage, K.; Kutok, J.; Feuerhake, F.; Kurtin, P.; Mihm, M.; Wu, B.; Pasqualucci, L.; & Others. (2005) “*Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response*.” *Blood*, 105(5):1851-1861.
- [46] Su, A.; Cooke, M.; Ching, K.; Hakak, Y.; Walker, J.; Wiltshire, T.; Orth, A.; Vega, R.; Sapinoso, L.; Moqrich, A. & Others (2002). *Large-scale analysis of the human and mouse transcriptomes*. *Proceedings of the National Academy of Sciences*, 99(7):4465-4470.
- [47] Evett, I. W., & Spiehler, E. J. (1987). “*Rule induction in forensic science*”. *KBS in Government*, Online Publications, páginas 107-118.

GiveMe Views: uma ferramenta de suporte a evolução de software baseada na análise de dados históricos

Alternate Title: GiveMe Views: a support tool to software evolution based on historical data analysis

Jacimar F. Tavares
Programa de Pós Graduação em
Ciência da Computação
Universidade Federal de Juiz de Fora
(UFJF)
jacimar.tavares@ice.ufjf.br

José Maria N. David
Programa de Pós Graduação em
Ciência da Computação
Universidade Federal de Juiz de Fora
(UFJF)
jose.david@ufjf.edu.br

Marco Antônio P. Araújo
Programa de Pós Graduação em
Ciência da Computação
Universidade Federal de Juiz de Fora
(UFJF) / IF Sudeste MG
marco.araujo@ice.ufjf.br

Regina Braga
Programa de Pós Graduação em
Ciência da Computação
Universidade Federal de Juiz de Fora
(UFJF)
regina.braga@ufjf.edu.br

Fernanda Campos
Programa de Pós Graduação em
Ciência da Computação
Universidade Federal de Juiz de Fora
(UFJF)
fernanda.campos@ufjf.edu.br

Glauco Carneiro
Programa de Pós Graduação em
Sistemas e Computação
Universidade Salvador (UNIFACS)
glauco.carneiro@unifacs.br

RESUMO

Este artigo apresenta a ferramenta *GiveMe Views*, capaz de indicar possíveis módulos e componentes que poderão ser impactados quando uma manutenção corretiva, adaptativa ou evolutiva é realizada. Fornece visualizações que permitem ao usuário obter informações sobre as relações existentes entre módulos e componentes baseando-se em análises estatísticas realizadas por um motor SAE (*Statistical Analysis Engine*) sobre dados históricos de software. Uma prova de conceitos foi realizada com o intuito de verificar a viabilidade da ferramenta em atividades de evolução de software em um contexto real de utilização. Ao final, os resultados da implantação da ferramenta em uma empresa de desenvolvimento de software são apresentados.

Palavras Chave

Evolução de Software, Visualização de Software.

ABSTRACT

This paper presents Give Me Views tool that can indicate possible modules and components that may be impacted when a corrective maintenance, adaptive or evolutionary is performed. Provides views that allow user to obtain information about the relationships between modules and components based on statistical analyzes performed by an SAE (*Statistical Analysis Engine*) on historical software data. A proof of concept was performed in order to verify the feasibility of the tool in software development activities in a real context of use. The tool deployment results in a software development company are presented.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.
Copyright SBC 2015.

Categories and Subject Descriptors

K.6.3 [Software Management]: *Software maintenance*.

General Terms

Management, Measurement, Design, Experimentation.

Keywords

Software Evolution, Software Visualization.

1. INTRODUÇÃO

Durante a evolução do software, alterações provenientes de manutenções realizadas podem ser observadas. Em [17] é definido que alterações no software constituem um ciclo de manutenção formado pelas fases:

- Requisição de mudanças: onde são definidas as mudanças a serem realizadas;
- Planejamento: se relaciona às atividades de (i) Compreensão do programa: na qual o software a ser alterado é analisado, buscando conhecimento sobre ele; e (ii) Análise do impacto: visa conhecer os riscos da efetivação da mudança, bem como o que será necessário alterar;
- Implementação da alteração: diz respeito à execução da alteração planejada. É formada pelas seguintes etapas: (i) Reestruturação para a mudança: contempla as alterações necessárias para reestruturar o sistema de modo a acomodar as novas implementações; (ii) Propagação da alteração: teoria fundamental neste trabalho, pois diz respeito a análise de componentes vizinhos a um componente alterado. Simplificando, cada vez que um componente é alterado, é dito que os componentes que possuem relação com ele (ou seja, vizinhos) devem ser verificados a fim de descobrir se alterações também serão necessárias neles. Alteração em um componente vizinho é denominada alteração secundária;
- Verificação e validação: verificação da alteração efetuada;

- **Re-documentação:** corresponde à adaptação da documentação existente ou à definição de nova documentação, objetivando registrar as alterações realizadas.

Este trabalho apresenta o relato do desenvolvimento da ferramenta *GiveMe Views*, um *plugin* para o ambiente de desenvolvimento *Eclipse*, que atua sobre projetos desenvolvidos em *Java*, visando apoiar a fase Implementação da alteração [17].

Criada para apoiar a análise visual de mudanças (apesar de não ser o objetivo deste artigo tratar sobre o tema visualização de software) realizadas em projetos de software da indústria, auxilia desenvolvedores nas tomadas de decisão sobre mudanças futuras. Os dados utilizados neste artigo, e para testes com o *GiveMe Views*, foram obtidos como resultado de parcerias estabelecidas com duas empresas de desenvolvimento de software de gestão empresarial. Uma delas utiliza um software próprio para a gestão de solicitações de mudança chamado *Service Center*. A outra utiliza o software *Mantis* [12]. Por questões de confidencialidade, a empresa que utiliza o *Service Center* será denominada neste artigo como empresa parceira 1. Já a que utiliza o *Mantis*, empresa parceira 2. Ambas as bases de dados dos citados softwares serão chamadas de repositórios de mudanças.

A motivação para o desenvolvimento da ferramenta surgiu a partir (i) da identificação de um problema no processo de manutenção da empresa parceira 1 e, (ii) uma particularidade encontrada no processo de registro e controle das solicitações de mudança, também na empresa parceira 1, ambos apresentados a seguir:

O problema: a empresa estava insatisfeita com o número de solicitações de mudanças reprovadas pelo setor de qualidade. As reprovações se davam (i) pela falta de controle sobre a evolução do software e (ii) a complexidade do projeto de código dos produtos. Ao longo dos anos e dos vários ciclos de manutenção, os produtos se afastaram de suas arquiteturas iniciais dado que não havia um controle sobre a evolução. Alterações eram realizadas sem o devido planejamento da evolução do projeto de código. Isso fez com que os projetos de códigos dos softwares se tornassem confusos e de difícil manutenção. Mediante isso, desenvolvedores que realizam as solicitações de mudanças têm dificuldade para identificar, por exemplo, quais componentes devem ser alterados quando outros componentes também forem alterados. Isso acontecia pois eles não possuíam nenhum tipo de solução (técnica, metodologia, ferramenta, *framework*, ou modelo) que fornecesse esse tipo de indicação.

A particularidade: está relacionada às informações de rastreabilidade que são adicionadas sempre que uma solicitação de mudança era finalizada, ou seja: módulos e componentes alterados. O termo módulo é designado para classificar: (i) softwares (produtos que a empresa comercializa), (ii) módulos (como módulo financeiro, módulo fiscal, entre outros) e (iii) pacotes de código (*packages*), tais como os criados em ambientes como *Netbeans* e *Eclipse*. Já o termo componente refere-se a: (i) classes ou arquivos de código alterados, (ii) formulários (tais como *jFrame* e *jDialog*), (iii) *DLLs* (bibliotecas de código e dados) [7] e (iv) *scripts* de banco de dados. A rastreabilidade é dita como as ligações existentes entre módulos e componentes que sofreram algum tipo de manutenção ao longo da resolução da solicitação de mudança. Permite que *links* sejam criados e mantidos entre artefatos gerados e alterados durante o ciclo de vida de um software [13]. Os módulos e componentes alterados são obtidos através dos sistemas de controle de versão que, tal como dito em [14], podem fornecer informações úteis para entender a evolução do software.

A empresa parceira 2 possuía a mesma dificuldade da empresa parceira 1 para estimar módulos e componentes que poderiam ser alterados sempre que um outro dado módulo ou componente fosse alterado. Em contrapartida, não possuía informações de rastreabilidade associadas às suas solicitações de mudança.

As informações de rastreabilidade, adicionadas às respectivas solicitações de mudanças, na empresa parceira 1, não eram usadas para ajudar na tomada de decisões futuras, mas somente para registro do que foi alterado pelo desenvolvedor, caso necessária uma consulta futura.

O problema a ser tratado neste artigo é a falta de uma ferramenta que apoie a análise da evolução de software utilizando dados históricos de solicitações de mudanças e informações de rastreabilidade de software e que atue como um *plugin* para o ambiente *Eclipse*. A principal contribuição está na definição de uma ferramenta livre que possibilite que desenvolvedores possam estimar os impactos que alterações no código fonte podem ocasionar, considerando para isso, dados históricos de todas as versões de um software analisado e cálculos estatísticos automaticamente gerados pela ferramenta, diferentemente de outras ferramentas de mesmo propósito que levam em consideração somente análises realizadas sobre o projeto de código, estaticamente e sobre uma única versão.

1.1 Metodologia

Inicialmente foi utilizado o *framework GiveMe Metrics* [9], parte inicial desta pesquisa, para conduzir a extração dos dados históricos de solicitações de mudanças dos repositórios de mudanças das empresas parceiras 1 e 2, respectivamente. Com os dados em mãos, foi conduzido um processo de análise junto ao Departamento de Estatística da Universidade Federal de Juiz de Fora (reunindo pesquisadores das áreas de Estatística e Ciência da Computação). Esse processo permitiu a implementação de um motor estatístico, chamado SAE - *Statistical Analysis Engine* (subseção 3.3). O SAE possibilitou ao *GiveMe Views* indicar quais as chances estatísticas de um dado módulo, ou componente, ser impactado quando outro módulo ou componente fosse alterado. Posteriormente, foi desenvolvida a ferramenta *GiveMe Views* que foi avaliada junto a empresa parceira 1. O objetivo deste artigo é apresentar a ferramenta *GiveMe Views*, bem como sua atuação na resolução do problema encontrado nas empresas parceiras 1 e 2.

Este artigo está dividido da seguinte forma: a Seção 2 apresenta os trabalhos relacionados. A Seção 3 introduz os princípios e o projeto da *GiveMe Views*. A Seção 4 apresenta a avaliação da primeira versão da ferramenta, realizada através de uma prova de conceitos. A seção 5 apresenta o cenário atual do *GiveMe Views* e suas principais funcionalidades implementadas. A seção 6 apresenta as considerações finais, os resultados obtidos e os trabalhos futuros.

2. TRABALHOS RELACIONADOS

Ferramentas para apoiar o processo de manutenção de software têm sido propostas na literatura técnica com frequência. Nesse contexto, algumas tratam de técnicas de visualização [3] como uma forma de tornar esse processo mais efetivo ao apresentar uma abordagem semiestruturada que permite a geração de visões gráficas, em forma de diagramas de classes, que foram diretamente geradas a partir da análise do projeto de código fonte de um software. O objetivo é obter a visão geral da arquitetura e dos relacionamentos entre as entidades envolvidas. Entretanto, os

dados históricos de software não são utilizados para o entendimento das relações entre as entidades, através apenas da geração de visões. Já em [4] é argumentado que, devido às mudanças que um software sofre durante seu ciclo de vida, é comum que, em alguns casos, suas arquiteturas se afastem do projeto original e até mesmo se tornem degradadas. Nesse sentido, é apresentada a ferramenta *Buch* que permite a geração de visualizações em forma de grafos a partir da arquitetura existente, utilizando, para isso, técnicas de agrupamento. Os grafos de relacionamento gerados por *Buch*, no entanto, não dão significância estatística à relação existente entre as entidades, se atendo apenas em apresentá-los.

SQuAVisIT [16] é um *framework* constituído por ferramentas e *plugins* que se conectam a repositórios e geram diferentes tipos de visualizações mediante a realização de análises. Essa proposta se difere das apresentadas em outros trabalhos relacionados por permitir a análise de código de diferentes linguagens. Entretanto, repositórios de dados históricos não são analisados no intuito de se obter informações sobre os relacionamentos entre as entidades em um contexto histórico.

3. GIVEME VIEWS: PRINCÍPIOS E PROJETO

A premissa para utilização do *GiveMe Views* é possuir dados históricos de software que possam ser extraídos de três diferentes tipos de repositórios. São eles: (i) de código fonte, (ii) de solicitação de mudanças e (iii) de processos de desenvolvimento de software. Por enquanto, *GiveMe Views* é capaz de analisar apenas dados provenientes dos repositórios (i) e (ii), como por exemplo dados de solicitações de mudanças e dados de *logs* de repositórios de código fonte.

Para que os dados pudessem ser extraídos das empresas parceiras, foi utilizado o *GiveMe Metrics Framework* [9]. Trata-se de um *framework* conceitual para extração de métricas e dados históricos sobre evolução de software em três diferentes tipos de repositórios.

A Figura 1 apresenta uma visão geral sobre a proposta de pesquisa, exibindo onde a ferramenta *GiveMe Views* atua. Mostra as parcerias realizadas (Empresas parceiras 1 e 2) que forneceram acesso aos seus repositórios de mudanças que foram analisados. O *framework* *GiveMe Metrics* foi utilizado e o conjunto de dados de saída da extração foram armazenados no repositório de dados históricos da ferramenta *GiveMe Views*, denominado *GiveMe Repository*.

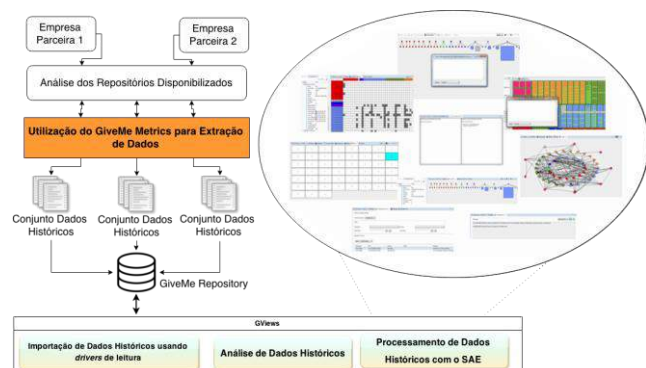


Figura 1. Visão geral da solução

O conjunto de dados de saída é a entrada para a *GiveMe Views*, que o analisa, processa e mantém as informações resultantes para

serem acessadas pelas visualizações, como ilustrado na elipse da figura.

Para que os dados mantidos no *GiveMe Repository* possam ser manipulados pelo *GiveMe Views*, foi necessária a implementação de *drivers* de leitura (subseção 3.2) que atuam na importação.

Os dados históricos, extraídos das empresas, foram analisados por um conjunto de pesquisadores. Através desta análise observou-se que os dados referentes a uma solicitação de mudança continham informações de manipulação de um ou mais componentes, no caso dos dados da empresa parceira 1. Também foi observado que um dado componente do software era sempre modificado quando outro componente também era modificado. Isso pode ser um indício de um acoplamento entre eles, de forma que alterações em um poderiam impactar diretamente no outro [15]. Observando essa questão, foi desenvolvido o motor SAE (subseção 3.3) e posteriormente o *GiveMe Views*.

3.1 Drivers de Leitura

Um importante conceito implementado no *GiveMe Views* é o de *driver*. São implementações que permitem a manipulação dos dados históricos independente do formato, que está restrito aos suportados pelo *GiveMe Metrics*: *.xml*, *.txt*, *.html*, *.csv* e *.xls*. Um conjunto de dados históricos pode ser formado, por exemplo, por um arquivo *.csv* e um *.txt*, ou somente um *.xls*.

Um *driver* é implementado para ler conjuntos de dados exportados por cada uma das ferramentas que compõem o *framework* *GiveMe Metrics*. Na versão atual do *GiveMe Views*, há suporte para leitura de dados exportados pela ferramenta Mantis [12] (como é o caso da empresa parceira 2), e dados exportados do repositório customizado da empresa parceira 1, o *Service Center*. É importante destacar que *GiveMe Views* suporta o desenvolvimento de *drivers* para leitura de dados históricos provenientes de outras fontes, e não somente para um conjunto restrito de ferramentas suportadas pelo *framework* *GiveMe Metrics*.

3.2 GiveMe Repository

É um repositório de dados históricos, e também o nome do *plugin* desenvolvido para automatizar a sua construção. A função do *plugin* é eliminar a responsabilidade do usuário na criação da estrutura de pastas do repositório. Além disso, implementa toda a lógica de manipulação do *GiveMe Repository*. A Figura 2 ilustra a estrutura do repositório *GiveMe Repository* gerada pelo *plugin*.

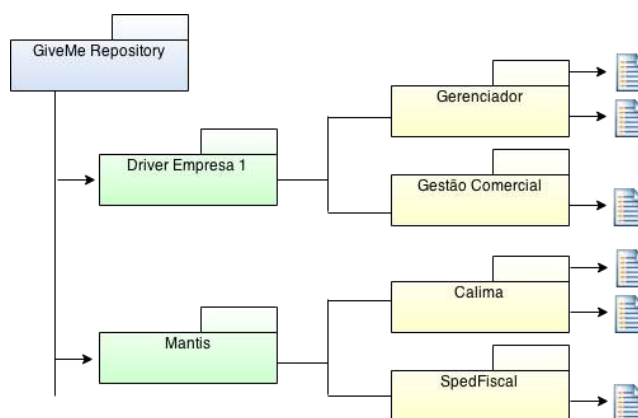


Figura 2. Estrutura do repositório gerado pelo *plugin* *GiveMe Repository*

Conforme a Figura 2 o espaço para armazenamento de dados para dois *drivers* diferentes são destacados: o *Driver Empresa 1* e o *driver* desenvolvido para a ferramenta *Mantis*.

Dentro de cada espaço, há outros espaços que são específicos para cada projeto que se tenha dados históricos para analisar. No espaço para o *Driver Empresa 1* há dados históricos de dois projetos: Gerenciador e Gestão Comercial. No espaço *Mantis* há dados históricos de outros dois projetos: Calima e *SpedFiscal*. Todos os quatro projetos mencionados são reais e foram disponibilizados através da parceria com as empresas 1 e 2, respectivamente.

Para gerar a estrutura do *GiveMe Repository*, o usuário deve apenas informar o caminho físico de sua escolha para que ele seja criado. Como resultado, todas as vezes que *GiveMe Views* necessitar analisar um conjunto de dados históricos, pertencentes a um projeto, o repositório será localizado.

3.3 SAE - Statistical Analysis Engine

O motor SAE foi concebido basicamente a partir de uma questão a ser investigada: historicamente, qual o percentual em que um dado componente foi modificado sempre que outro componente também foi modificado? Para responder esta questão foram consideradas teorias estatísticas de Distribuição de Frequência Simples e Conjunta [5]. A Figura 3 ilustra a definição do SAE.

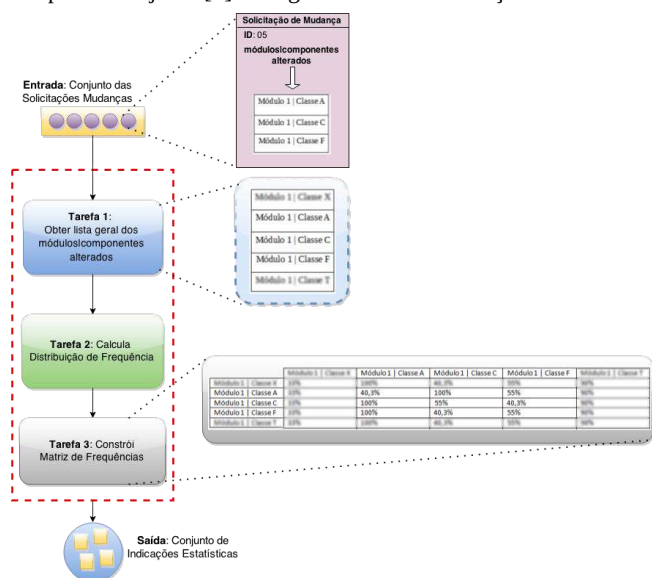


Figura 3. Sequência de tarefas do SAE

Os módulos da Figura 3 que estão dentro da linha pontilhada representam o SAE. Como entrada, tem-se um conjunto de solicitações de mudanças. Cada solicitação é única, e possui um conjunto de informações, como classes e módulos alterados para resolução daquela solicitação de mudança, como pode ser visto no *zoom* aplicado na solicitação de mudança de ID 05.

A primeira tarefa do SAE é obter a lista geral dos módulos e componentes alterados, sem repetições entre todos os conjuntos de informações. Em seguida, são executados os cálculos estatísticos (Tarefa 2) e construída a matriz que possui as frequências calculadas (Tarefa 3).

O *zoom* na Tarefa 3 mostra um exemplo de matriz com as estatísticas obtidas. Como saída do SAE, tem-se um conjunto de

estatísticas sobre qualquer método de um software, através do qual é possível avaliar, por exemplo, o impacto de uma alteração em um dado método que se deseja dar manutenção.

Para o caso da empresa parceira 2, (e de outras empresas que, porventura, não tenham informações de rastreabilidade associadas diretamente às solicitações de mudança), a ferramenta *GiveMe Views* prevê a leitura de outro conjunto de dados como entrada, que são arquivos de *log* de histórico de *commits* obtidos de repositórios de código fonte. Atualmente há suporte para repositórios SVN (*SubVersion*) [10] e GIT [11].

Após o desenvolvimento do SAE, foi construída a ferramenta *GiveMe Views*, a qual implementará o SAE em um módulo específico, com suas respectivas visualizações. A subseção seguinte apresenta a arquitetura do *GiveMe Views*.

3.4 Arquitetura do GiveMe Views

A arquitetura foi projetada com dois objetivos principais: (i) permitir a execução do processamento estatístico dos dados históricos ao invocar o motor SAE, e (ii) apoiar a persistência de informações originadas do processamento estatístico com o SAE, permitindo que elas sejam acessadas por visualizações ou por outras ferramentas, posteriormente integradas à *GiveMe Views*. A Figura 4 apresenta a arquitetura, e a relação entre os módulos implementados.

Uma entrada para a ferramenta *GiveMe Views* é a indicação de um módulo/componente que se deseja dar manutenção. Para isto, basta selecionar o componente que a ferramenta capturará o seu nome e o módulo ao qual ele pertence, concatenando essas informações separadas pelo caracter *pipe* “|”, conforme ilustrado na Figura 5.

Em seguida, os dados históricos referentes ao projeto alvo da manutenção são importados do *GiveMe Repository* e filtrados, para a análise estatística, bem como para o cálculo das métricas disponíveis. A informação gerada é então persistida, disponibilizada para exportação ou para acesso de visualizações e de ferramentas, via provedor. Para executar a ferramenta *GiveMe Views*, basta selecionar o projeto que se deseja dar manutenção e, com o botão direito do mouse, acessar a opção *Visualizar with GiveMe Views*.

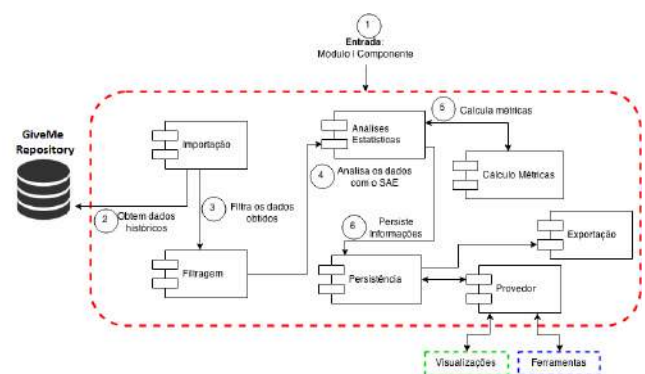


Figura 4. Arquitetura *GiveMe Views*

- Importação: responsável por selecionar no repositório o conjunto de dados históricos referentes ao projeto em manutenção;
- Filtragem: realiza a filtragem dos dados importados, desconsiderando, por exemplo, dados de solicitações de mudanças que não foram concluídas e, portanto, não afetaram

nenhum trecho de código do projeto em manutenção. Portanto, são irrelevantes no processamento estatístico;

- Análises Estatísticas: responsável por implementar o SAE. Além disso, é responsável por formatar as informações nos formatos necessários para os diferentes tipos de visualização disponíveis no GiveMe Views;
- Cálculo Métricas: calcula as métricas sobre dados históricos que são disponíveis no GiveMe Views. A lista completa das métricas está disponível em [2].
- Persistência: módulo responsável por manter em memória todas as informações estatísticas processadas, bem como as métricas e as informações gerenciais da ferramenta;
- Provedor: único ponto pelo qual visualizações e ferramentas a serem integradas ao GiveMe Views podem obter informações sobre os dados históricos;
- Exportação: módulo responsável por implementar os relatórios disponíveis para exportação.

4. AVALIAÇÃO DA FERRAMENTA

Esta seção descreve uma prova de conceitos com o objetivo de analisar a viabilidade da ferramenta GiveMe Views para apoiar a avaliação de impacto em atividades de manutenção de software. O objetivo é apresentado segundo a abordagem

Goal/Question/Metric (GQM) [1]: “**Analisar** a ferramenta GiveMe Views **com a finalidade de** verificar a viabilidade de uso na tarefa de identificação dos possíveis módulos e componentes que serão impactados quando uma dada manutenção corretiva, adaptativa ou evolutiva for efetuada em um componente de um módulo **com respeito** a alterações no projeto de **código do ponto de vista do** desenvolvedor que realizará a manutenção **no contexto de** evolução de software”.

A prova de conceito contou com a participação do gerente de projetos da empresa parceira 1. Após a utilização do GiveMe Metrics [9] para extrair os dados do repositório de solicitação de mudanças, a ferramenta GiveMe Views foi apresentada por meio de um treinamento realizado em um único dia. Posteriormente, o gerente analisou o repositório de mudanças e selecionou uma que

envolvia uma manutenção corretiva já concluída. O objetivo é usar o GiveMe Views para verificar a sua capacidade de indicar os mesmos módulos e componentes alterados na solicitação de mudança selecionada. A solicitação selecionada foi a manutenção corretiva realizada no Componente *clsCalculoNota*, pertencente ao módulo *DLL – Calculos*, uma vez que se tratava de uma solicitação concluída recentemente pelo próprio gerente de projetos. Nesse contexto, as seguintes hipóteses foram definidas:

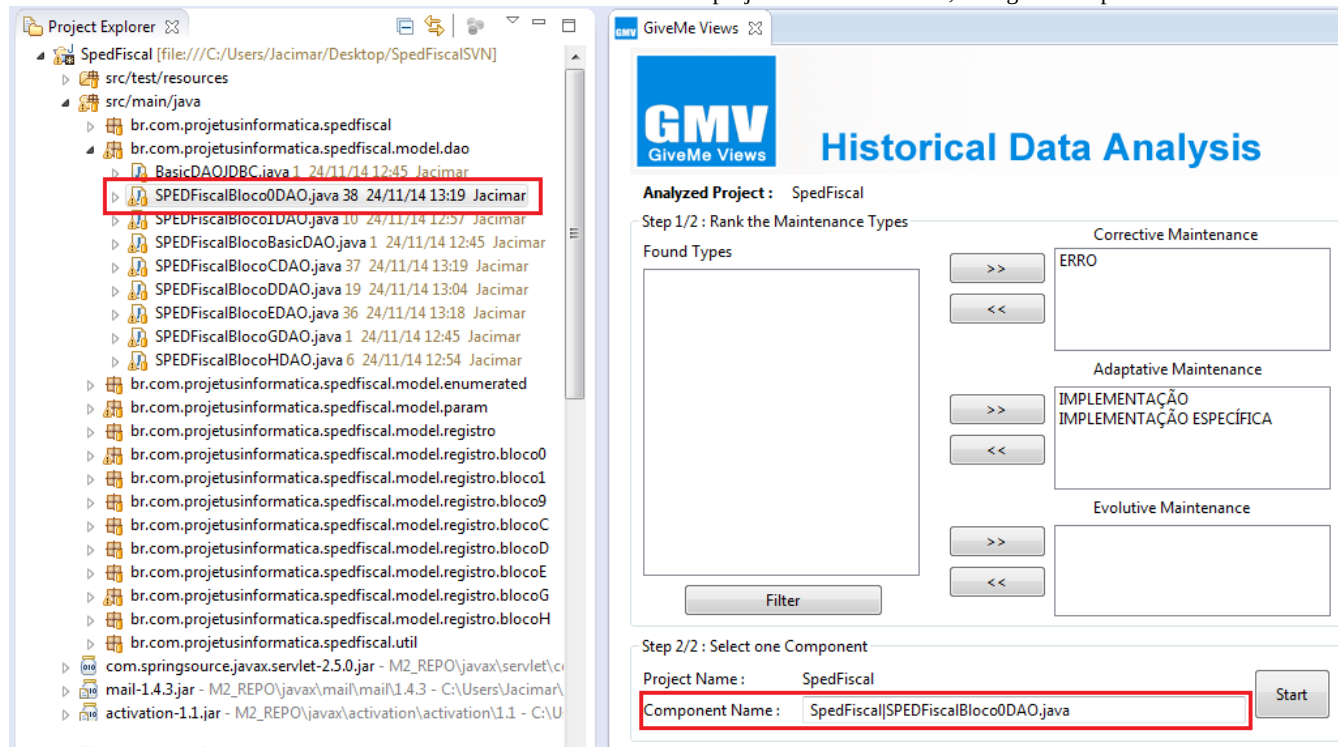


Figura 5. Tela principal do GiveMe Views e a seleção de um componente para verificação do impacto de mudança

- H_{nula} : a ferramenta GiveMe Views não é capaz de apoiar a identificação de quais componentes e módulos poderão ser impactados quando uma dada manutenção corretiva for realizada no componente *clsCalculoNota*;
- $H_{alternativa}$: a ferramenta GiveMe Views é capaz de apoiar a identificação de quais componentes e módulos poderão ser impactados quando uma dada manutenção corretiva for realizada no componente *clsCalculoNota*.

A Figura 6 mostra a visualização Tree View [6] gerada a partir da execução da ferramenta GiveMe Views, considerando a

necessidade de verificar quais componentes poderão ser impactados quando uma manutenção for realizada no componente *clsCalculoNota*, pertencente ao módulo *DLL – Cálculos*.

A porcentagem entre parênteses é o resultado do cálculo estatístico realizado pela ferramenta GiveMe Views. Assim, *fNotaEntrada*, componente pertencente ao módulo Gerenciador, foi o componente com maiores chances estatísticas de ser impactado quando uma manutenção corretiva for realizada no componente *clsCalculoNota*, pertencente ao módulo *DLL – Cálculos* (12,12% de chance).

Ainda, através da Figura 6, é possível perceber que a ferramenta pode indicar impactos não somente em componentes de um mesmo módulo, mas também entre outros módulos que são interligados. Como exemplo, há mais chances estatísticas do componente *fNotaEntrada*, pertencente ao Módulo *Gerenciador*, ser impactado do que o componente *mdlTipos* (10,60%) pertencente ao mesmo módulo que se deseja alterar.

A prova de conceito apresentou evidências iniciais de que a ferramenta pode apoiar a identificação de componentes que de fato foram modificados, no caso selecionado pelo gerente de projetos da empresa parceira, aceitando-se a hipótese alternativa da questão de pesquisa. Entretanto, avaliações adicionais devem ser realizadas no sentido de evidenciar situações nas quais um único experimento não permitiu revelar.

Como ameaças à validade do experimento podemos citar o fato da avaliação ter se baseado em uma manutenção realizada pela empresa parceira 1. Caso ela não tenha impactado os componentes de fato necessários, a indicação dada pelo *GiveMe Views* poderia não corresponder à realidade.

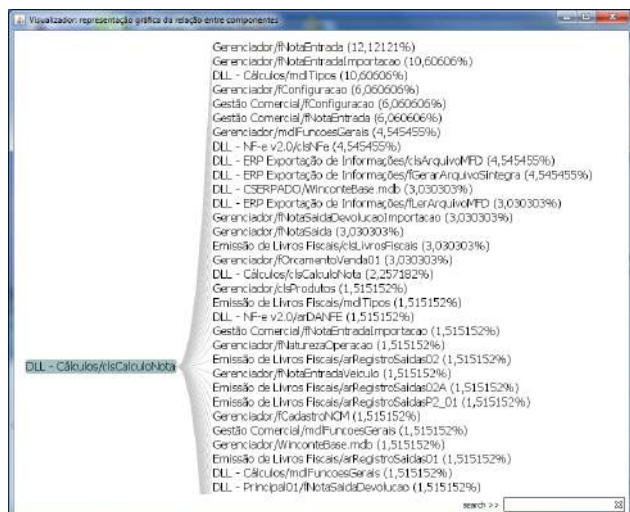


Figura 6. Visualização Tree View

5. GIVEME VIEWS: CENÁRIO ATUAL

Após a condução da prova de conceitos, foram realizadas implementações e adaptações nas funcionalidades do *GiveMe Views* visando adequar a ferramenta a uma nova necessidade: fornecer indicações de métodos que poderão ser impactados quando um dado método de uma classe fosse alterado, e não somente indicar classes ou arquivos de código, formulários (tais como *jFrame* e *jDialog*), DLLs (bibliotecas de código e dados) e scripts de banco de dados, que podem ser impactados quando uma dada manutenção fosse realizada.

A adaptação se tornou necessária porque, em alguns projetos que serão alterados, indicar apenas as classes que poderão ser impactadas com mudanças em classes relacionadas pode não ser suficiente, por exemplo. A existência de classes, como algumas encontradas em projetos na empresa parceira 1, que possuem mais de 15.000 linhas de código evidenciaram esta necessidade. Indicar que uma classe dessas possa ser impactada se uma alteração for realizada em outra dada classe pode ser uma informação muito vaga. As novas implementações e adaptações realizadas visaram apenas integrar a ferramenta *GiveMe Views* com a ferramenta *GiveMe Trace*, apresentada na seção 5.1.

5.1 Integração GiveMe Views e GiveMe Trace

GiveMe Trace [15] é uma ferramenta de apoio a rastreabilidade de software. Trata-se de um *plugin* para o ambiente *Eclipse*. Atua gerando dados históricos sobre a rastreabilidade de software ao se conectar automaticamente a repositórios SVN e GIT. É capaz de reportar informações em dois formatos diferentes, basicamente:

- *Arrays de strings*: dado um número de *commit* (nome dado ao envio alterações de código ao repositório de código fonte) ou um intervalo de *commits* (ou seja, um conjunto de números que representam diferentes *commits* realizados para um projeto), *GiveMe Trace* retorna uma lista dos métodos alterados por *commit*. Cada *string* retornada segue o mesmo padrão: nomeDaClasse.java|nomeDoMetodo (nome da classe é concatenado com um *pipe* e com o nome do método alterado). Um *array* retornado contém um conjunto de métodos de todas as classes de código fonte que foram modificadas num dado *commit*, ou um intervalo deles.
- Arquivos texto (txt): dado um número ou intervalo de *commits*, *GiveMe Trace* retorna um arquivo de *log* de alterações realizadas no repositório de código fonte, idêntico ao fornecido pelo próprio gerenciador de código fonte. A diferença consiste na granularidade das informações. A Figura 7, extraída de [15], mostra um exemplo de arquivo de *log* gerado na análise dos *commits* de número 19 e 20 (informados pelo usuário da ferramenta). Nela é possível ver que todas as modificações realizadas estão registradas no nível de métodos alterados, separados das classes as quais pertencem por um *pipe*.

As informações geradas pela ferramenta *GiveMe Trace* não podem ser geradas nativamente pelos sistemas de controle de versões de código fonte (SVN e GIT). Através dessas informações a ferramenta *GiveMe Views* passou a realizar o processamento estatístico com o SAE, considerando um nível de granularidade mais baixa, ou seja, no nível de método.

```
Commit: 19
Autor: USUARIO
Date: Tue Nov 11 13:35:37 BRST 2014

Mensagem: final da implementação dos Ranges;
Resolvi warnings:

Fechamento do projeto:

M /givemetrace/src/givemetrace/implementations/TraceSit.java|preCoreToLogMethod
M /givemetrace/src/givemetrace/implementations/TraceSit.java|preCoreToLogClass
M /givemetrace/src/givemetrace/implementations/TraceSit.java|preCoreToArrayMethod
M /givemetrace/src/givemetrace/implementations/TraceSit.java|preCoreToArrayClass
M /givemetrace/src/givemetrace/implementations/TraceSit.java|getCountCommit

Commit: 20
Autor: USUARIO
Date: Tue Nov 18 18:25:43 BRST 2014

Mensagem: tratamento dos url nos métodos do provider para evitar exception por
não reconhecer que o url é um repositório

M /givemetrace/src/givemetrace/views/vpGiveMeTraceMain.java|createPartControl
```

Figura 7. Trecho do arquivo de rastreabilidade

A seção 5.2 apresenta o conjunto de recursos e visualizações desenvolvido para o *GiveMe Views*, já considerando a integração com a ferramenta *GiveMe Trace*.

5.2 Recursos e visualizações

Atualmente, *GiveMe Views* disponibiliza recursos e visualizações que visam apoiar desenvolvedores na tomada de decisão em atividades de evolução de software. Nesta subseção serão apresentadas as três principais visualizações disponíveis. Para

executar o *GiveMe Views*, foram utilizados dados históricos do projeto *SpedFiscal*, cedido pela empresa parceira 2, que é um sistema para apoiar a gestão empresarial.

A Figura 8 mostra a *Tree View* criada após a seleção do método *getRegistro0200*, (pertencente à classe *SPEDFiscalBloco0DAO.java*), para análise de impacto, bem com os métodos que se relacionam com ele e as chances estatísticas, historicamente falando, de sofrerem impacto com uma manutenção.

Além de exibir os métodos que podem ser impactados quando o método selecionado for modificado, a visualização *Tree View* ainda exibe um número que representa a porcentagem de chances de isso ocorrer considerando o histórico de alterações, calculado pelo SAE.

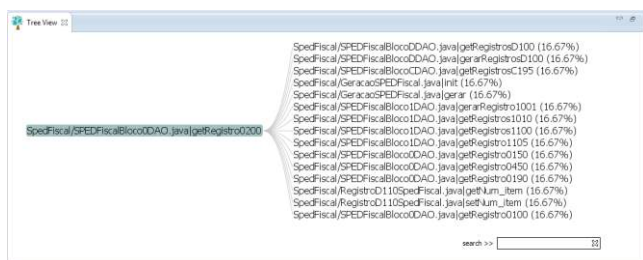


Figura 8. Visualização *Tree View*

Assim como a *Tree View*, a visualização *Deep View* também exibe os métodos que podem ser impactados quando um dado método for alterado e a porcentagem estatística de um impactar o outro, historicamente, de forma ordenada pela maior porcentagem estatística. Entretanto, *Deep View* possui um diferencial: é capaz de analisar os impactos em diferentes níveis de profundidade. A Figura 9 mostra os métodos que podem ser impactados caso uma alteração seja feita no método *getRegistro0200*, pertencente a classe *SPEDFiscalBloco0DAO*. Numa atividade de manutenção, o desenvolvedor que modificasse o método *getRegistro0200* (método pai), teria que verificar, em todos os métodos que se relacionam com ele (filhos), se alguma alteração será necessária.

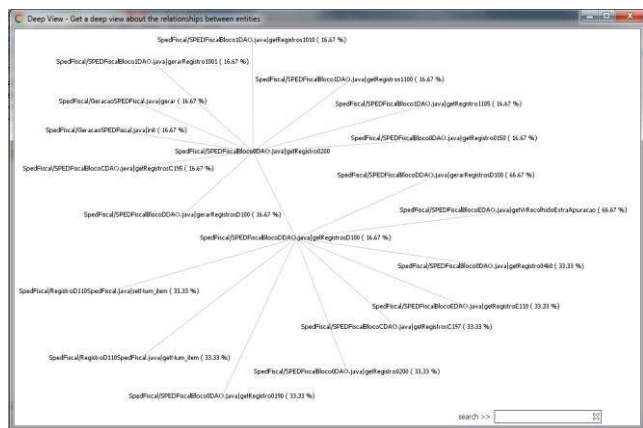


Figura 9. Visualização *Deep View*

Caso um dos filhos seja alterado, um novo ciclo se iniciará, sendo necessário verificar se os filhos do filho modificado foram impactados. Para analisar os métodos que podem ser impactados quando um método filho for alterado, basta clicar com o botão direito do mouse sobre um método que a visualização *Deep View* automaticamente irá exibir os filhos.

Durante a realização de uma atividade de manutenção, o desenvolvedor pode se deparar com questões de tempo que o impeça de verificar todos os métodos indicados pelo *GiveMe Views* como possíveis alvos de manutenção. Nesse contexto, o que se pode fazer é considerar aqueles com maiores chances estatísticas, e começar por eles. Além das porcentagens estatísticas, tem-se outro recurso que pode ser considerado, que está disponível na visualização *Modules and Components View* (Figura 10). A primeira informação que se tem é que o projeto *SpedFiscal* foi alvo de 468 manutenções, sendo que delas, 6 afetaram o componente *getRegistro0200*.

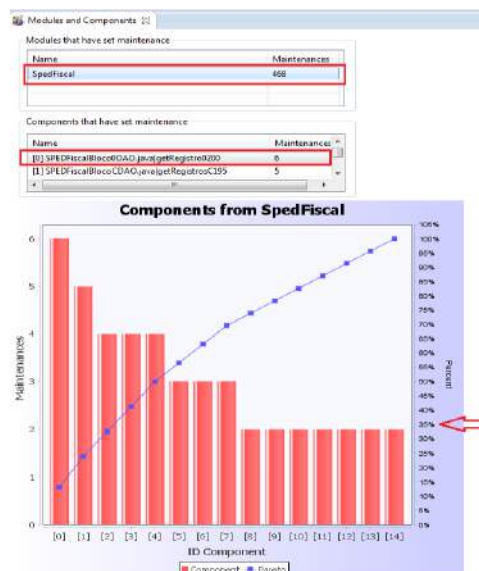


Figura 10. Visualização *Modules and Components View*

O gráfico exibido à direita mostra o número de manutenções por componente e uma curva de Pareto, que indica que 35% das manutenções ocorreram nos métodos representados pelos ids 0, 1 e 2, tornando-os boas opções de verificação de manutenção inicial.

5.3 Resultados Obtidos

Como resultados da implantação da ferramenta na empresa parceira 1 tem-se:

Cálculo de métricas: um dos resultados deste trabalho é que *GiveMe Views* fornece um meio para calcular um conjunto de métricas, cujos valores antes eram desconhecidos, proporcionando acesso a informações que podem auxiliar no processo de tomada de decisão, por exemplo: foi possível à empresa parceira 1 perceber que o número de desdobramentos (reaberturas de solicitações de mudança) abertos é maior que o número de solicitações de mudanças. Isso, na prática, indica que a empresa gasta mais recursos para solucionar problemas que surgem após a modificação de um componente se comparado aos recursos gastos com solicitações provenientes de usuários. Isso pode mostrar que o nível de acoplamento do sistema compromete a manutenção, indicando a necessidade de efetuar melhorias no projeto dos seus produtos.

Atualização da equipe de suporte: a métrica Porcentagem Acerto de Estimativa de um Módulo permitiu à empresa descobrir que o setor de suporte não estava sendo eficiente na tarefa de identificação do trecho que deverá ser alterado, mediante a

solicitação de mudança do cliente. Assim que um chamado é aberto, a equipe de suporte cadastra o possível módulo e o componente que deverá ser alterado. *GiveMe Views* mostrou que, em três anos de manutenções, apenas 35% das vezes essa informação foi corretamente cadastrada. Nesse caso, a equipe de suporte necessitará passar por novo treinamento para conhecer a frequência de modificação dos produtos ao longo do tempo. Cabe ressaltar que, tal métrica pode não se aplicar a outros contextos de outras empresas, dado a necessidade de se possuir dados históricos que permitam o cálculo estatístico.

Apoio ao setor de qualidade: o processo de mudanças da empresa parceira 1 consiste na abertura de uma solicitação de mudança, na sua execução e verificação da mesma. A etapa de verificação é de responsabilidade do setor de qualidade. Ao receber uma solicitação de mudança para a correção de um defeito, por exemplo, deve-se, além de verificar se o defeito foi corretamente solucionado, verificar se novos defeitos não foram inseridos em outros componentes acoplados. Com *GiveMe Views*, o setor de qualidade passou a ter uma referência que indica os pontos que devem ser verificados, facilitando assim o trabalho antes feito manualmente e baseado na experiência de profissionais do setor de qualidade.

Percepção de uso da ferramenta: inicialmente, *GiveMe Views* foi testada pelo gerente de projetos da empresa parceira 1, que assimilou de imediato suas funcionalidades e utilização. Foi reportado que a ferramenta é de fato útil na identificação de possíveis pontos que poderão ser impactados dada uma solicitação de mudança, e que seu uso ajudará a reduzir o tempo atualmente gasto na identificação de possíveis pontos que serão impactados, bem como a avaliação do que foi alterado.

6. CONSIDERAÇÕES FINAIS

O problema abordado neste artigo não se aplica apenas ao contexto das empresas parceiras 1 e 2. De uma forma geral, empresas que não possuem documentação atualizada da arquitetura de seus produtos, bem como informações sobre seu projeto de código, podem ser potenciais usuárias da ferramenta *GiveMe Views*. Como contribuições tem-se o apoio prestado pela ferramenta *GiveMe Views* à manutenção de software, dado que ela proporciona uma visão gráfica e geral sobre os relacionamentos entre módulos e componentes, sem que seja necessário acesso ao código fonte, pois a ferramenta leva em consideração somente dados históricos em suas análises.

As limitações da ferramenta *GiveMe Views* são aquelas impostas pelo conjunto de dados históricos que são analisados. Caso o conjunto de dados de entrada seja ruim, as análises podem não ser precisas e as indicações podem ser irrelevantes. Um conjunto de dados é dito ruim se as modificações realizadas no código fonte de um software em manutenções anteriores não tiverem impactado métodos que realmente têm algum tipo de relação. Nesse caso, *GiveMe Views* indicaria pontos a serem alterados que podem não ter relação de fato com o componente a ser alterado, mesmo que com chances estatísticas pequenas de impacto.

Como trabalhos futuros, *GiveMe Views* será integrada à ferramenta *SourceMiner* [8] que provê a visão dos relacionamentos considerando o contexto do código fonte fornecendo, portanto, a visão do código no momento atual do software, isto é, considerando sua versão mais recente. A integração proporcionará ao usuário uma troca de contexto em tempo de execução, permitindo analisar a relação entre módulos e

componentes do ponto de vista do código fonte e do ponto de vista histórico, baseado em estatística, como o fornecido pelo *GiveMe Views*. Através desta integração busca-se avaliar a forma pela qual as atividades de compreensão podem auxiliar na Evolução de Software.

7. AGRADECIMENTOS

À CAPES, à FAPEMIG e ao CNPq pelo apoio financeiro, ao professor José Antônio Teodoro, do Departamento de Estatística da Universidade Federal de Juiz de Fora e às empresas parceiras.

8. REFERÊNCIAS

- [1] Basili, Victor R. Caldiera, Gianluigi H. Rombach, Dieter. (1994). "The Goal Question Metric Approach". Chapter in Encyclopedia of Software Engineering, Wiley, 1994.
- [2] *GiveMe Views*: <http://givemeinfra.com.br/givemeviews>. Acessado em jun de 2014.
- [3] Haitzer, Thomas. Uwe Zdun. "Semi-automated architectural abstraction specifications for supporting software evolution". Science of Computer Programming, Vol. 90, Part B, 2014.
- [4] Mancoridis, S.; Mitchell, B.S.; Chen, Y.; Gansner, E.R. "Bunch: a clustering tool for the recovery and maintenance of software system structures," (ICSM '99) Proc. IEEE International Conference on Software Maintenance, 1999.
- [5] Meyer, Paul L. "Probabilidade – Aplicações à Estatística". 2ª ed. LTC – Livros Técnicos e Científicos Editora Ltda. 1993.
- [6] Prefuse: <http://www.prefuse.org/>. Acessado em jan de 2014.
- [7] Microsoft DLL: <http://support.microsoft.com/kb/815065/pt-br>. Acessado em 06 fev de 2015..
- [8] Carneiro, G. F. "SourceMiner: um ambiente integrado para Visualização multi-perspectiva de software". 230 f. Tese (doutorado) – Universidade Federal da Bahia, Inst. Matemática, Doutorado em Ciência da Computação. 2011.
- [9] Tavares, J., David, J. M. N., Araújo, M. A.P., Braga, R., Campos, F. C. A. (2014), "GiveMe Metrics – Um framework conceitual para extração de dados históricos sobre a evolução do software". X Simpósio Brasileiro de Sistemas de Informação (SBSI), Londrina/PR.
- [10] SUBVERSION: <https://subversion.apache.org/> Acessado em 10 de fev de 2015.
- [11] GIT: <http://git-scm.com/>. Acessado em 10 de fev de 2015.
- [12] Mantis Bugtracker: <https://www.mantisbt.org/>. Acessado em 06 de fev de 2015.
- [13] Walters, B.; Shaffer, T.; Sharif, b; Kagdi, H. Capturing software traceability links from developers' eye gazes. Proc. 22nd Int. Conf. on Program Comprehension. ACM, 2014.
- [14] Mohan, K.; Xu, P.; Cao, L.; Ramesh, B. Improving change management in software development: Integrating traceability and software configuration management. Decision Support Systems, v.45, n.4, p. 922 - 936, 2008.
- [15] Lélis, C. A. S, Araújo, M. P, David, J. M. N. *GiveMe Trace*: Uma ferramenta para apoio a rastreabilidade de software. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) - Universidade Federal de Juiz de Fora, 2014.
- [16] Van Den Brand, M.; Roubtsov, S.; Serebrenik, A. (2009). "SQuAVisiT: A Flexible Tool for Visual Software Analytics," Software Maintenance and Reengineering, 2009. CSMR '09. 13th European Conference on, pp. 331- 332. GIT: <http://git-scm.com/>. Acessado em 10 de fev de 2015.
- [17] Sneed, H.M. "Software evolution: A roadmap". Software Maintenance. Proc. IEEE International Conference, 2001.

Aplicações Android de Realidade Aumentada em Arquitetura Extensível, Flexível e Adaptável

Alternative Title: Android Augmented Reality Applications in Extensible, Flexible, and Adaptable Architecture

Tiago Araújo

Programa de Pós-Graduação em
Ciência da Computação
Universidade Federal do Pará
tiagodavi70@gmail.com

Nikolas Carneiro

Programa de Pós-Graduação em
Ciência da Computação
Universidade Federal do Pará
nikolas.carneiro@gmail.com

Brunelli Miranda

Faculdade de Computação
Universidade Federal do Pará
brunelli.miranda@gmail.com

Carlos Santos

Programa de Pós-Graduação em
Ciência da Computação
Universidade Federal do Pará
gustavo.cbcc@gmail.com

Bianchi Meiguins

Programa de Pós-Graduação em
Ciência da Computação
Universidade Federal do Pará
bianchi@ufpa.br

RESUMO

A proposta deste trabalho é apresentar uma arquitetura para aplicações Android de Realidade Aumentada Móvel (RAM) que seja extensível, flexível e adaptável. Extensível por permitir a incorporação de novas funcionalidades na aplicação; flexível por permitir a troca de conteúdo/dados gerando aplicações personalizadas; e adaptável aos diversos tamanhos de telas que os dispositivos móveis apresentam. Para isso, foi considerada uma arquitetura baseada no padrão MVC (Model, View, Controller) adaptado para contexto da plataforma móvel Android, nesta adaptação a lógica de negócio sai da camada de controle e vai para camada de visão, sendo assim, a camada de controle fica responsável apenas por gerenciar as requisições entre as camadas modelo e visão, garantindo uma modularização dessa aplicação. Também foi utilizado o padrão de interface de usuário Fragments, visando uma melhor adaptação aos diversos tamanhos de telas dos dispositivos móveis. Além disso, foi aplicado o padrão de projeto Proxy Remoto, para abstração da origem dos dados (local ou remota), e o padrão de projeto Facade para facilitar a realização de consultas e filtros nos dados. Por fim, são apresentados cenários de uso com três tipos diferentes de dados e dispositivos com tamanhos de telas diferentes para validação da arquitetura proposta.

Palavras-Chave

Realidade Aumentada, Aplicações Móveis, Arquitetura.

ABSTRACT

This work aims to present an architecture for Android Mobile Augmented Reality (MAR) applications which has to be

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.
Copyright SBC 2015.

extensible, flexible and adaptable. Extensible to allow the addition of new functionalities in the application; flexible to allow the change of content/data generating custom applications; and adaptable to the many mobile devices screen sizes. For this reason, we considered an architecture based on the MVC pattern adapted to the context of Android mobile platform, in this adaptation the business logic get out the controller layer and reaches the view layer, in so doing, the controller layer becomes responsible only for managing the request between the model and view layers, ensuring the application modularization. Also, the Fragments user interface pattern was utilized, aiming to a better adaptation to the many mobile devices screen sizes. Furthermore, we applied the Remote Proxy design pattern, for abstraction of data source (local or remote), and the Facade design pattern to facilitate the use of queries and filters in data. Finally, we present usage scenarios with different data and screen size devices to validate the proposed architecture.

Categories and Subject Descriptors

D.2.11 [Software Engineering]: Software Architectures – Domain-specific architectures, Patterns.

General Terms

Management, Design.

Keywords

Augmented Reality, Mobile Applications, Architecture.

1. INTRODUÇÃO

A evolução tecnológica dos smartphones e tablets, principalmente no que diz respeito a qualidade de renderização e câmera, uso do GPS e sensores, como bússola e acelerômetro, viabilizou a utilização de aplicações de Realidade Aumentada nestes dispositivos. As áreas de aplicações da Realidade Aumentada Móvel (RAM) são as mais variadas possíveis, como as áreas de jogos, viagens, compras on-line, ciências, saúde, redes sociais, entre outras [11].

Com todos esses recursos tecnológicos disponíveis o desafio passa a ser o desenvolvimento de aplicações que utilizem esses recursos de forma integrada, seja em aplicações nativas ou híbridas (aplicações web para dispositivos móveis). Outras questões relevantes consideram as restrições e características diferentes de hardware e software dos vários dispositivos, questões de segurança, variabilidades na interface gráfica do usuário, o consumo de energia do dispositivo, etc [12].

Apesar do crescimento acelerado desta área e de inúmeras aplicações disponíveis para uso, não há padrões de desenvolvimento definidos para aplicações RAM, a maioria das aplicações apresenta pouca flexibilidade de dados não permitindo seu uso em outros domínios, há necessidade de evolução na precisão das tecnologias de localização e reconhecimento de padrões, e pouco espaço para apresentação das informações virtuais [8].

Este trabalho propõe um modelo arquitetural para aplicações de Realidade Aumentada Móvel, voltado para plataforma Android, que permite a extensibilidade da aplicação com baixo custo de implementação, adaptação a uma diversidade de dispositivos com tamanhos diferentes de tela, e flexibilidade na obtenção e utilização dos dados.

Neste sentido, foi utilizada uma adaptação do padrão MVC para contexto de aplicações de RA em plataforma móvel, onde a lógica de negócio sai da camada de controle e vai para camada de visão, e a camada de controle fica responsável por gerenciar as requisições entre a camada de modelo e camada visão.

Também foi utilizado o padrão de interface de usuário Fragments, visando a facilidade à adaptação, gerenciamento e responsividade aos diversos tamanhos de telas dos dispositivos. Além disso, foi aplicado o padrão de projeto Proxy Remoto, para abstração da origem dos dados (local ou remota), e o padrão de projeto Facade para facilitar a realização de consultas e filtros nos dados. Também foram seguidas as boas práticas e principais diretrizes para o desenvolvimento de aplicações móveis e aplicações de RA.

Por fim, serão apresentados cenários de uso com dados e dispositivos com tamanho de telas diferentes para validação da arquitetura proposta. O primeiro cenário considera um roteiro artístico das obras da Universidade Federal do Pará, o segundo considera um roteiro educacional no Museu Emílio Goeldi e o terceiro considera um roteiro turístico com atrações turísticas da cidade de Belém-Pará-Brasil. As aplicações móveis têm como principais funcionalidades: guiar os usuários em ambiente aumentado por GPS, guiar o usuário em mapas 2D, e apresentar conteúdos aumentados através de reconhecimento de padrões QR Code. O ambiente aumentado é apresentado através da tela do dispositivo combinando a cena real com conteúdos virtuais (imagem, texto, vídeo, objetos 2D/3D).

2. REALIDADE AUMENTADA MÓVEL

A Realidade Aumentada Móvel (RAM) é a aplicação do conceito de RA em plataformas móveis, que significa misturar conteúdos virtuais e cenas do ambiente real na tela do dispositivo móvel [6].

Na RAM a cena do mundo real é capturada pela câmera do dispositivo móvel e associada a uma camada de conteúdo para apresentação na tela do dispositivo, a aplicação cria outra camada com conteúdos virtuais, e então sobrepõe uma camada à outra para apresentação ao usuário.

Normalmente, as aplicações de RAM apresentam duas abordagens para misturar conteúdo virtual nas cenas reais: marcadores ou localização. A abordagem de marcadores utiliza um software de reconhecimento de um padrão em particular (como QR Code) para selecionar o conteúdo, e a abordagem por localização faz uso do GPS do dispositivo e da orientação da câmera para definir o que deve ser apresentado ao usuário.

Como principais características de aplicações RAM, têm-se [7]:

- Apresentação do ambiente aumentado na tela dos dispositivos móveis;
- Tecnologia de rastreamento: GPS, e reconhecimento de padrões em imagens;
- Sistema gráfico é responsável pela renderização dos objetos virtuais em cenas aumentadas;
- Sistema de mistura de mundos: faz a mistura da cena real com objetos virtuais.

Os desafios no desenvolvimento de aplicações RAM são [8, 12]:

- Integração com sensores dos dispositivos,
- Baixa precisão das tecnologias de rastreamento
- Restrições e diferentes características físicas dos dispositivos;
- Variabilidade na interface do usuário,
- Falta de padrões comuns que de fato sejam adotados pelos desenvolvedores de aplicativos RAM;
- Consumo de energia

Algumas diretrizes devem ser consideradas para o desenvolvimento de aplicações RAM, são elas [2, 3]:

- Levar em consideração o perfil do usuário alvo, o uso outdoor, o uso de uma ou duas mãos, e o tempo de uso da aplicação;
- Seguir os bons princípios de usabilidade para RA e aplicações móveis: cena aumentada limpa, ícones e textos grandes e em camadas, interação com objetos 3D, e considerar o mundo real;
- Levar em consideração as restrições dos dispositivos: reflexão da tela e a falta de precisão no rastreamento;
- Percepção e a cognição do usuário devem ser estimuladas;
- Apresentação das informações virtuais: considerar a quantidade de informação, representação da informação, local das informações, utilização de múltiplas visões;
- Avaliação: adaptar diretrizes de avaliação de usabilidade para aplicações móveis e RA, avaliar um domínio por vez, e utilizar usuários reais.

3. ARGUIDE

3.1 Aspectos Conceituais

ARguide é uma aplicação de realidade aumentada móvel que pode receber vários tipos de conteúdos georeferenciados e fornecer aos usuários este conteúdo de forma aumentada. Este aplicativo projetado para utilizar os recursos disponíveis em dispositivos móveis com sistema operacional Android, como: sensores, câmera, bússola, GPS, acesso a Internet, etc e tornar o conteúdo

da aplicação acessível e explorável por meio da realidade aumentada.

A ideia desta aplicação é guiar o usuário que deseja conhecer mais sobre um determinado conteúdo, estimulando a visita de lugares e pontos de interesses, com a utilização de recursos como mapa, navegador de realidade aumentada, leitores de QR Code, e arquivos multimídia.

A aplicação foi arquitetada de modo a oferecer os recursos de uma aplicação RAM para o contexto personalizado do fornecedor de conteúdo da aplicação, sem a necessidade de programação para a plataforma Android, entretanto nesta versão da aplicação é necessário algum conhecimento em configurações de aplicações Android. Ou seja, nesta versão da aplicação o usuário insere os dados (contexto personalizado) em um projeto genérico do ARguide e gera o arquivo executável.

A Figura 1 apresenta o fluxo de construção de uma aplicação ARguide com contexto personalizável. O primeiro passo trata-se dos dados georeferenciados fornecidos por uma pessoa ou organização, neste caso pode ser um hospital, um governo, uma entidade, uma empresa etc, que deseja fornecer um conteúdo aumentado para seus usuários. No segundo passo o conteúdo é armazenado em diretórios específicos dentro de um projeto genérico do ARguide e é gerado um executável que pode ser instalado em dispositivos móveis com o sistema operacional Android. No terceiro passo o fornecedor do conteúdo terá uma aplicação de realidade aumentada móvel com os seus dados fornecidos na primeira etapa dispondo dos recursos de realidade aumentada móvel para visita, exploração e descoberta desse conteúdo.

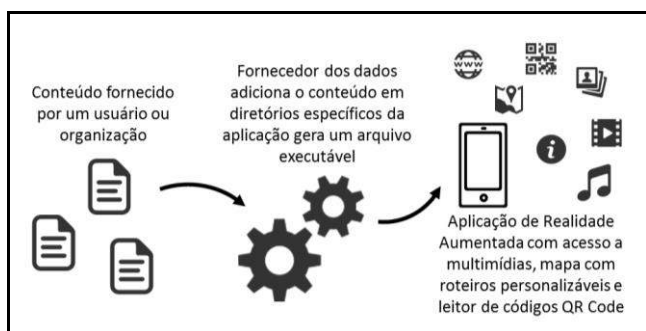


Figura 1. Ideia do funcionamento da aplicação para gerar aplicações personalizadas.

A aplicação gerada, independente do conteúdo adicionado, possui três partes, são elas: aba mapa, onde podem ser visualizados POIs (Pontos de Interesse do inglês *Points of Interest*) e rotas em um mapa 2D; aba navegador de RA, que permite visualizar pontos de interesse com recursos de GPS, bússola e acelerômetro; e um leitor QR Code, que permite recuperar conteúdos associados a QR Codes.

3.2 Arquitetura

O ARguide foi projetado para oferecer os recursos de uma aplicação RAM independente dos dados que forem fornecidos. Desse modo, para garantir que o ARguide ofereça tais recursos é necessário garantir uma boa modularidade de código entre os componentes que realizam as trocas de dados na aplicação, desde o nível arquitetural.

Na arquitetura do ARguide decidiu-se utilizar uma abordagem semelhante ao padrão MVC (*Model, View, Controller*), que pode ser traduzido como Modelo, Visão e Controlador), porém com algumas modificações que fornecem uma melhor adequação ao estilo de programação de aplicações Android com a linguagem Java.

Esta decisão arquitetural provocou uma ligeira, porém efetiva modificação no papel de cada componente do padrão MVC. Esta modificação considera reduzir a quantidade de lógica aplicada no controlador e passar essa lógica para a visão, dando assim mais autonomia para as visões e uma menor dependência do controlador. A Figura 2 ilustra como essa ideia foi aplicada no padrão MVC.

Devido a essa alteração no comportamento do componente Controlador decidiu-se alterar o nome seu nome para Roteador (Router), uma vez que este componente não possui mais lógica fazendo apenas o papel de encontrar quais modelos devem ser alterados, quando isso for requisitado. Outro papel importante associado ao Roteador é a separação entre os componentes Visão e Modelo permitindo que alterações em um desses componentes não reflitam no outro.

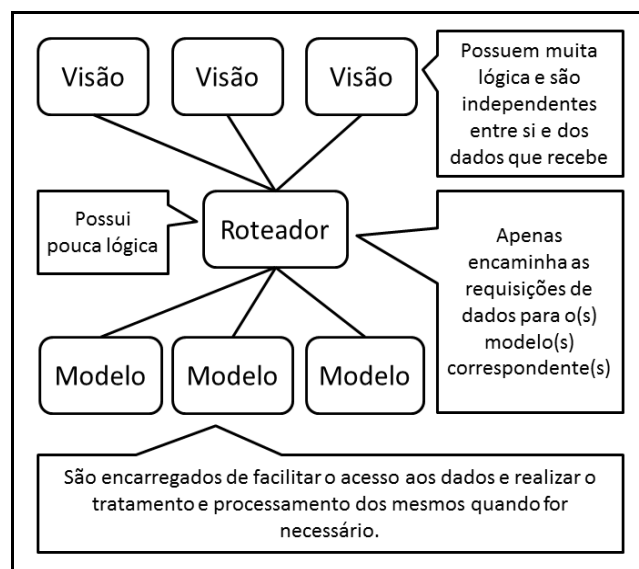


Figura 2. Modificação proposta no padrão MVC.

Aplicações Android possuem como *entry point*¹ uma classe Activity, que não só define o que está sendo visto na tela (por meio das classes View² e de layouts em XML), mas também é responsável pela manutenção do contexto da aplicação e pelo ciclo de vida de uma janela na aplicação. Sendo assim, muita responsabilidade e lógica da aplicação estão diretamente associadas com a classe Activity, logo essa modificação realizada no padrão MVC adequa melhor esta arquitetura ao estilo de programação de aplicações Android com Java.

No padrão MVC o Controlador possui lógica para as ações realizadas pelo usuário na Visão, ou seja, cada interação do usuário na interface gráfica dispara uma chamada para o Controlador que sabe o que realizar para cada tipo de interação do

¹ *Entry Point*: primeiro arquivo a ser executado em uma aplicação.

² *View*: Classe da API Android responsável pelo desenho elementos gráficos na tela do usuário.

usuário. Com a modificação proposta o Controlador deixar e ter essa lógica e passa apenas a direcionar chamadas para os respectivos Modelos quando for solicitado. Dessa maneira as Visões passam a ter mais lógica e apenas vão disparar chamadas para o Roteador/Controlador quando necessitarem realizar alguma operação de modificação de dados. Essa medida permite adicionar e remover novas Visões sem a necessidade de modificações no Roteador ou no Modelo.

Uma das inspirações para a modificação proposta foi o padrão MVP que também é uma variante do padrão MVC, mas é aplicado para construções de interfaces gráficas com o usuário. O padrão MVP define o *entry point* na Visão, que considera a mesma como uma estrutura de Widgets³ que devem ser responsivos às interações do usuário, o Apresentador (*Presenter*) como responsável por mediar a comunicação entre o Modelo (*Model*) e a Visão (*View*), e o Modelo, trata dos dados e a lógica de acesso aos dados [10]. Uma das formas conhecidas do MVP é o MVP *Passive View* [4] que canaliza todas as interações e os retornos do Modelo para as Visões pelo Apresentador, mas mantendo parte da autonomia presente nas Visões. Entretanto neste padrão é criada uma instância de Apresentador para cada Visão deixando o projeto menos expansível, uma vez que para adicionar uma funcionalidade nova é necessário criar uma nova Visão e seu respectivo Apresentador.

Além dessa arquitetura neste projeto foi utilizado o padrão de interfaces gráficas de usuário Fragmentos (*Fragments*) [9], possibilitando criar interfaces mais ricas, interativas e ubíquas. Este padrão foi utilizado para melhor organizar as funcionalidades disponíveis para o usuário, assim como, permitir uma melhor coordenação entre as visões.

Cada Fragmento comporta-se dentro de seu próprio escopo, sendo capaz de tratar suas próprias interações, realizar processos independentes e gerenciar seus próprios componentes gráficos. Essa característica atribui mais flexibilidade e extensibilidade para a aplicação, que pode sofrer alterações, como remoção do mapa ou adição de uma nova forma de interagir, sem alterar o funcionamento dos outros meios de interação. Além disso, o padrão Fragmentos permite o design de layouts flexíveis e responsivos em aplicações móveis, o que significa que a mesma aplicação pode se adaptar a diferentes tamanhos de tela de dispositivos, evitando retrabalho de codificação para os diferentes dispositivos.

A arquitetura proposta é apresentada de maneira geral na Figura 3, que mostra o fluxo de interações do usuário e respectivas respostas do sistema passando entre módulos e submódulos verticalmente. Os módulos apresentados estão descritos a seguir:

- **Visão:** camada responsável por manipular os eventos da interface do usuário, apresentar os dados vindos da camada Modelo e encaminhar as interações do usuário para a camada Roteador, apenas quando for necessário realizar alterações em dados. Possui três sub-módulos intercambiáveis:
 - o **Mapa:** responsável por apresentar um mapa 2D, pontos de interesses, e rotas;
 - o **Navegador de RA:** responsável por apresentar conteúdo aumentado no dispositivo, tendo como

referência o GPS e a orientação da câmera do dispositivo (capturada através da bússola e acelerômetro);

- o **Leitor de QR Code:** responsável pela leitura de um marcador tipo QR Code, e apresentar conteúdo virtual aumentado do marcador na tela do dispositivo associado ao QR Code específico.
- **Roteador:** responsável pela comunicação entre o modelo e a visão, permitindo um intercâmbio fracamente dependente entre as duas partes. Essa característica permite que alterações realizadas na visão não acarretem alterações no modelo e o mesmo vale para alterações no modelo.
- **Modelo:** responsável pela definição dos dados a serem apresentados na Visão. É composto por:
 - o **Facilitador (Facade):** auxilia a realização de consultas e modificações mais facilmente, este é o padrão de projetos Facade descrito por Gamma e colaboradores [5];
 - o **Proxy Modelo:** Abstrai as requisições aos dados. Para seus clientes não importa de onde vieram os dados, este padrão se encarrega de delegar a função de recuperação de dados para um de seus sub-módulos, este módulo trata-se do padrão de projetos Proxy Remoto [5].
 - **Modelo Local:** quando os dados requisitados estão armazenados no dispositivo do usuário este módulo recupera os dados e passa para a camada Roteador;
 - **POIs Proxy:** quando os dados não estão armazenados no dispositivo do usuário este módulo recupera os dados através da rede, armazena os dados na memória local do dispositivo.

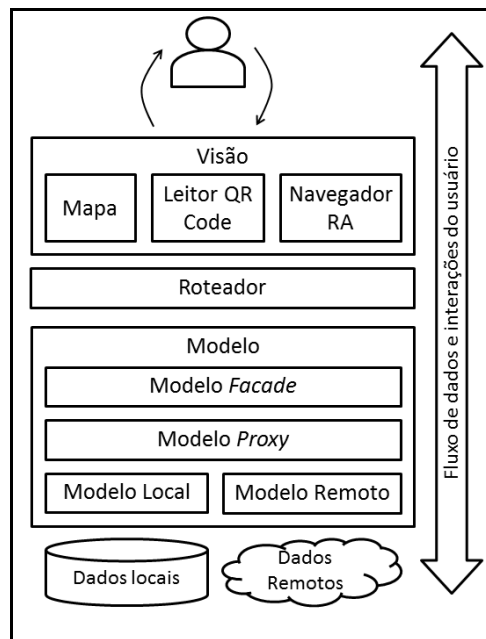


Figura 3. Arquitetura proposta e o fluxo de dados e interações do usuário

³ Widgets: Componentes de interface gráfica do usuário. Exemplo: Botões, check-boxs, caixas de texto, etc.

3.3 Funcionalidades

O ARguide possui resumidamente três tipos de interações, que são: interações com Mapa, Navegador RA (Realidade Aumentada) e QR Code. As interações com Mapa e Navegador RA foram divididas em abas na aplicação e as interações com leitor de QR Code são acionadas através de um botão.

Com a arquitetura proposta torna-se menos custoso adicionar ou retirar uma funcionalidade da aplicação. Como os dados da aplicação podem variar de acordo com o fornecedor, pode ser que o mesmo deseje adicionar novas funcionalidades e não queira uma funcionalidade já existente. Como as funcionalidade estão associadas a camada de Visão, basta remover ou adicionar uma nova Visão para ter esse efeito. Embora, nesta versão seja necessário algumas habilidades de programação para remover e adicionar funcionalidade esta tarefa foi simplificada com a arquitetura proposta.

Abaixo é descrita cada uma dessas funcionalidades e o que o usuário pode realizar por meio dessas funcionalidades:

3.3.1 Aba Mapa

A aba Mapa contém um mapa desenvolvido com a tecnologia Google Maps, que permite aos usuários a realização de interações em dispositivos móveis que viabilizam a visualização, busca e seleção de POIs em um mapa 2D. A partir dessa aba o usuário pode realizar as seguintes operações:

- **Zoom e Pan:** são operações de ampliar e mover o mapa na tela do dispositivo, estas operações são importantes tanto para dar uma visão geral dos POIs, assim como, visualizar detalhes de um ou alguns POIs.
- **Selecionar Rota:** uma vez que dados inseridos na aplicação são georeferenciados, o aplicativo é capaz de traçar uma rota de visitação POIs de acordo com a posição atual do usuário.
- **Selecionar POI:** com os pontos de interesse marcados no mapa o usuário pode clicar/tocar em um POI realizando uma seleção. Quando um POI é selecionado é apresentado ao usuário um conjunto de conteúdos multimídia sobre esse POI ao usuário.
- **Acompanhar sua localização em relação a Rota:** trata-se do usuário poder acompanhar sua trajetória em relação a rota que foi sugerida pela aplicação. Como a localização atual do usuário é mostrada no mapa ele pode saber o que já foi visitado e o que ainda falta visitar.

3.3.2 Aba Navegador RA

A aba Navegador RA apresenta ao usuário o stream de sua câmera (vídeo), representando o mundo real, acrescido de marcadores virtuais que são posicionados em frente aos POIs, marcando a posição do POI quando o usuário aponta sua câmera para este. Com essa marcação o usuário tem sua percepção aumentada em relação a esse POI, pois sabe quais POIs possuem conteúdo na aplicação e qual a direção exata de cada ponto de interesse. Esta é uma abordagem semelhante à *Magic Lens*, descrita por Cawood [1], onde a posição de cada POI é dada pela informação de geolocalização do mesmo.

A partir dessa aba o usuário pode realizar as seguintes operações:

- **Visualizar Posição/Direção do POI:** permite ao usuário encontrar POIs baseado na sua localização atual, obtida a

partir do GPS do dispositivo móvel, ao direcionar sua câmera para os POIs disponíveis na aplicação.

- **Selecionar POI:** permite que o usuário da aplicação clique/toque no ícone virtual posicionado na direção do POI selecionando o mesmo. Ao selecionar um POI o usuário tem acesso aos seus conteúdos multimídia.
- **Ajustar Raio Alcance:** permite que o usuário configure um raio de interesse para a exibição dos ícones virtuais. Este raio é medido em quilômetros e tem como centro a posição atual do usuário. Todos os POIs que estiverem a uma distância maior que esse raio não serão visualizados pelo usuário. Esta operação pode ser utilizada para que o usuário possa se concentrar na busca por elementos mais próximos a ele.
- **Visualizar Radar:** a aplicação disponibiliza um pequeno radar no canto superior direito da tela. Esse radar mostra um ponto para cada POI localizado dentro do raio de interesse configurado. Isto permite que POIs que não estejam sendo visualizados, por não estarem na direção da câmera do usuário, possam ser notados.

3.3.3 Aba Leitor de QR Code

A aba de leitor de QR Code busca e executa um serviço de leitura de códigos QR Code disponível no dispositivo do usuário. Este serviço realiza a leitura dos códigos extraíndo as informações contidas em tais marcadores. Ao receber as informações extraídas pelo serviço a aplicação pode disponibilizar o acesso à lista de informações do ponto de interesse em questão (se o POI for identificado) ou mostrar uma mídia diretamente.

3.3.4 Funcionalidades Gerais

As funcionalidades gerais da aplicação estão sempre disponíveis para o usuário independente de qual aba esteja ativa. Estas funcionalidades objetivam auxiliar o usuário em atividades gerais e em momentos de dúvida. As funcionalidades gerais estão descritas abaixo:

- **Visualizar Conteúdo Multimídia:** Esta operação está sempre disponível para o usuário quando ele seleciona um POI em qualquer uma das abas, ela permite que o usuário visualize arquivos relacionados com o POI selecionado, por exemplo, vídeos, áudios, textos, imagens, e website. Estes dados multimídia ajudam os usuários a conhecerem mais sobre o POI durante, antes ou depois da visitação ao mesmo.
- **Limpar Seleção:** esta opção limpa a seleção realizada pelo usuário em um momento anterior. Esta opção não é obrigatória para realizar outra seleção, mas remove da tela elementos que podem não interessar mais para o usuário.
- **Ajuda:** auxilia o usuário dando dicas de utilização da aplicação. A ajuda oferece uma descrição completa da utilização da aplicação e suas funcionalidades.

3.4 Geração de Aplicações Personalizadas

Com ARguide é possível gerar aplicações de realidade aumentada com dados personalizados a partir de um projeto (códigos fonte da aplicação organizados em um conjunto de pastas) genérico, ou seja sem os dados. Neste caso o fornecedor dos dados deve copiar seus dados em um formato específico em uma pasta desse projeto. Um requisito para os dados fornecidos é que eles devem ser georeferenciados.

Tendo copiado os dados na pasta destinada a esse propósito, o projeto está pronto para ser compilado gerando um arquivo executável, neste caso um arquivo com extensão APK, que são arquivos instaladores de aplicações em dispositivos Android.

Após gerar o APK o fornecedor pode distribuir sua aplicação para seus usuários da forma que lhe for conveniente, uma das formas populares de distribuição de aplicações Android é através do mercado eletrônico Google Play.

4. CENÁRIOS DE USO

Os cenários de uso apresentados neste trabalho foram configurados para testar a utilização de diferentes tipos de dados, assim como, diferentes tamanhos de telas com a finalidade de validar as características da arquitetura proposta.

Foram realizados três testes para os dados e dois testes para os tamanhos de tela, sendo que esses foram separados em cenários de uso. Foram feitos configurados três cenários, são eles: cenário turístico, artístico, e educativo.

Cada um desses cenários usou seus respectivos dados, sendo, turísticos, artísticos e educativos, representado os testes de dados e dois dispositivos com o sistema operacional Android, um Tablet e um Smartphone, representando os testes de tamanhos de tela.

O Tablet utilizado nos testes é um Samsung Galaxy Tab 3 com as seguintes configurações: processador Dual-Core 1,5GHz, 1,5GB de memória RAM, e tela de 8 polegadas. E o Smartphone utilizado nos testes foi um Samsung Yound Duos TV com as seguintes configurações: processador de 1GHz, 750MB de memória RAM e tela de 3,2 polegadas.

Os cenários de uso da aplicação estão descritos nas subseções a seguir.

4.1 Cenário Artístico

O cenário artístico possui dados de obras artísticas da Universidade Federal do Pará. Com este cenário um usuário pode explorar e visitar as obras artísticas espalhadas pela universidade conhecendo suas localizações e mais sobre sua história, significados e autores.

Os possíveis *stakeholders*⁴ envolvidos neste cenário são, como fornecedor dos dados: organizações interessadas em divulgar as obras e artistas de uma determinada região, dando acesso aos significados, descrições e contexto histórico das obras. E como usuário: entusiastas, alunos de artes e pessoas interessadas em conhecer e visitar as obras artísticas em um determinado contexto.

A Figura 4 mostra um exemplo típico da utilização da aplicação em um cenário artístico. Neste o usuário está apontando seu Tablet para a obra artística Olho d'Água⁵ e está visualizando um marcador virtual na direção da localização exata da obra. Ao clicar neste marcador, o usuário pode ter acesso às mídias desta obra, como: textos explicativos, contexto histórico, resumo do autor, web site da obra, etc.

A Figura 5 mostra a captura de tela da aplicação no cenário artístico, desta vez em um Smartphone. Com a captura de tela é

possível ver com mais detalhes o marcador virtual na direção da obra artística. Neste é possível observar a utilização do navegador RA em um dispositivo com um tamanho de tela bem menor em relação ao Tablet.



Figura 4. Foto da utilização da aplicação em um cenário artístico com a utilização de Tablet no modo Navegador RA.



Figura 5. Captura de tela da utilização da aplicação em um cenário artístico com a utilização do Smartphone no modo navegador RA.

⁴ *Stakeholders*: são as partes interessadas na aplicação, ou seja, pessoas ou organizações que tenham interesse ou sejam afetados pela aplicação.

⁵ A escultura “Olho d’água”, do artista plástico Acácio Sobral: <http://www.portal.ufpa.br/imprensa/noticia.php?cod=3051>

4.2 Cenário Educativo

O cenário educativo possui dados de informativos sobre plantas e animais, e obras do Museu Emílio Goeldi⁶, situado em Belém-Pará-Brasil. Com este cenário imagina-se uma visita educativa ao museu com acesso aos conteúdos multimídias de animais, plantas, e obras estimulando o aprendizado e o interesse por esse tipo de conteúdo.

Os possíveis *stakeholders* envolvidos neste cenário são, como fornecedor dos dados: organizações educacionais com interesse em disponibilizar uma ferramenta de auxílio ao aprendizado e ao interesse pelo conteúdo ensinado. E como usuário: alunos e professores que desejam aprender e ensinar de maneira prática através da visita aumentada de bosques, museus e zoológicos.

A Figura 6 mostra uma foto da utilização da aplicação em um cenário educacional com a utilização do *Smartphone* e usando o leitor de QR Code para saber mais informações sobre a Arara-Azul no museu Emílio Goeldi. A ideia desta foto é mostrar como a aplicação pode ser utilizada em excursões educativas em zoológicos, museus, planetários, bosques, entre outros.



Figura 6. Foto da utilização da aplicação em cenário educacional com a utilização do *Smartphone* no modo Leitor de QR Code.

A Figura 7 apresenta a visualização de conteúdo (neste caso texto, mas poderiam ser fotos, vídeos, áudios e websites) logo após a

seleção de um POI, que neste caso ocorreu através da leitura de um código QR Code, mas esta seleção poderia ocorrer pelo toque em um ícone virtual no Navegador RA ou toque em um marcador no mapa.



Figura 7. Foto da utilização da aplicação em cenário educacional com a utilização de *Tablet* para visualização do conteúdo de texto relacionado com a Arara-Azul.

4.3 Cenário Turístico

O cenário turístico possui dados relacionados com os pontos turísticos da cidade de Belém-Pará-Brasil. Com este cenário um usuário pode conhecer e visitar os pontos turísticos da cidade com acesso a roteiros turísticos aumentados.

Os possíveis *stakeholders* envolvidos neste cenário são, como fornecedor dos dados: organizações interessadas em promover o turismo da cidade de forma mais interativa e tecnológica, dando um acesso fácil, rápido e aumentado para os pontos turísticos de uma cidade. E como usuário, turistas interessados em visitar a cidade e ao mesmo tempo conhecer mais sobre a história e obras arquitetônicas desta.

As Figuras 8 e 9 mostram a utilização da aplicação em um cenário turístico com o recurso de mapas 2D em *Tablet* e *Smartphone* respectivamente.

Na Figura 8 o usuário selecionou um POI (neste caso uma atração turística) tocando em um marcador no mapa, e lhe foi apresentada uma imagem do local, uma nota de 0 a 5 (estrelas azuis na imagem), e o nome do local. Também foi apresentada uma rota que pode ser visualizada pela linha verde e é ordenada através de numerações nos marcadores.

A Figura 9 mostra a mesma rota, porém entretendo em um dispositivo com tela menor.

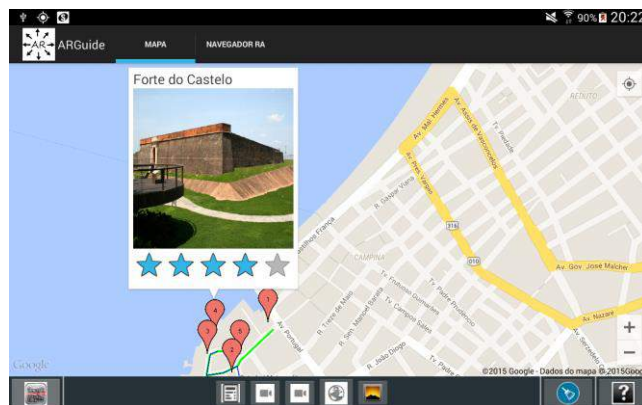


Figura 8. Captura de tela da aplicação em cenário turístico com a utilização de *Tablet* no modo mapa.

⁶ Museu Paraense Emílio Goeldi: <http://www.museu-goeldi.br/portal/>

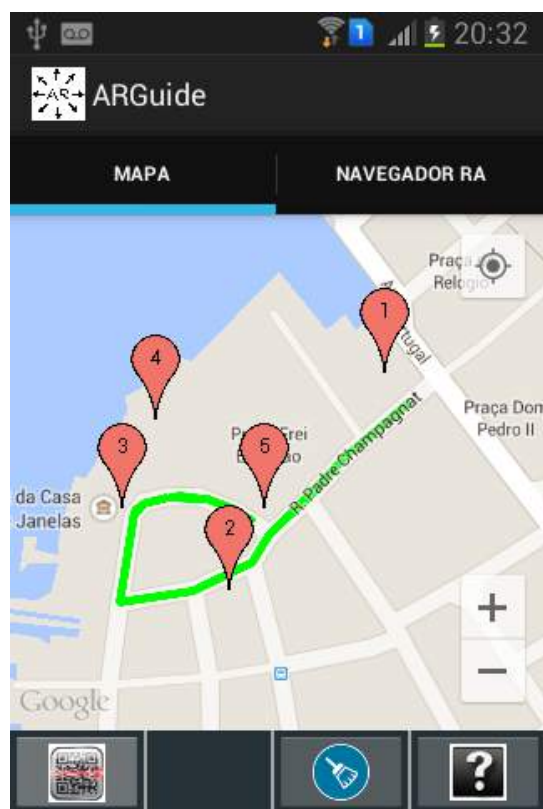


Figura 9. Captura de Tela da aplicação em cenário turístico com a utilização de Smartphone no modo mapa.

Embora cada um dos cenários tenha mostrado funcionalidades distintas é possível utilizar todas as funcionalidades em cada um desses cenários, assim como em outros cenários, dependendo apenas do conteúdo que será utilizado. Além disso, é possível adicionar e remover funcionalidades para melhor adequar a aplicação às necessidades do cenário de utilização. Por exemplo, em uma visita a um museu com muitos andares, pode ser útil incluir uma funcionalidade de reconhecimento de imagem, enquanto que a utilização do Navegador RA será prejudicada devido a ausência de rastreamento dos andares e a perda da precisão do GPS, dando a impressão que todos os POIs estão no mesmo andar.

5. CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma proposta de arquitetura para aplicações Android de Realidade Aumentada Móvel (RAM) extensível, flexível, e adaptável, sugerindo uma modificação no padrão MVC e considerando a utilização de padrões de projetos, como Proxy Remoto e Facade, assim como, a utilização do padrão de interfaces gráficas Fragments.

Dentre os pontos fortes da arquitetura proposta, pode-se destacar: a possibilidade de adicionar novas funcionalidades na aplicação devido a lógica da aplicação ter passado para as visões, basta adicionar uma nova visão que saiba se comunicar com o Router para ter uma nova funcionalidade; a possibilidade de adicionar vários tipos de dados, com o único requisito dos dados serem georeferenciados, permitindo que a aplicação seja genérica servindo a muitos propósitos; e a adaptação a diferentes tamanhos de tela que é uma característica muito comum em dispositivos móveis com o sistema operacional Android.

Dentre os pontos fracos da aplicação, pode-se destacar: a necessidade do fornecedor dos dados ter que lidar com a configuração do projeto (códigos e diretórios da aplicação) para gerar sua aplicação personalizada; a falta de teste de usabilidade para garantir uma boa experiência com o usuário; e a ausência de conteúdo 3D que é importante para uma experiência de realidade aumentada.

Os trabalhos futuros são aplicar melhorias nestes pontos fracos identificados com as seguintes medidas: desenvolver um gerador automático de aplicações, onde o fornecedor do conteúdo georeferenciado possa adicionar e organizar seus dados através de uma interface gráfica; realizar testes de usabilidade como checklists e experimentos com usuários reais; e adicionar mais funcionalidades como conteúdos 3D, reconhecimento de imagem e acesso a redes sociais como Facebook e Twitter.

6. REFERÊNCIAS

- [1] Cawood, S., Fiala, M. 2008. *Augmented Reality: A Practical Guide*. Pragmatic Bookshelf, (Dallas, Texas, USA, Jan, 2008).
- [2] de Sá, M., Churchill, E. 2012. Mobile Augmented Reality: A Design Perspective. *Human Factors in Augmented Reality Environments*. Springer, (New York, NY, USA), 139-164.
- [3] Dünser, A. et al. 2007. Applying HCI principles to AR systems design. *Mixed Reality User Interfaces: Specification, Authoring, Adaptation (MRUI'07) Workshop Proceedings*. (Charlotte, NC, USA), 37-42.
- [4] Fowler, M. 2002. *Patterns of Enterprise Application Architecture*. Addison-Wesley Longman Publishing Co. Inc, (Boston, MA, USA, Nov, 2002).
- [5] Gamma, E. et al. 1994. *Design Patterns: elements of reusable object-oriented software*. Addison-Wesley Longman Publishing Co., Inc, (Boston, MA, USA, Nov, 1994).
- [6] Grubert, J., Grasset, R. 2013. *Augmented Reality for Android Application Development*. Packt Publishing Ltd., (Birmingham, UK, Nov, 2013).
- [7] Joorabchi, M. et al, 2013. Real Challenges in Mobile App Development. *International Symposium on Empirical Software Engineering and Measurement. ACM / IEEE*, (Baltimore, Maryland, USA, Out 10-11, 2013), 15-24.
- [8] Martínez, H. et al. 2014. Drivers and Bottlenecks in the Adoption of Augmented Reality Applications. *Journal of Multimedia Theory and Application*. Vol. 2, No. 1, 27-44.
- [9] Nudelman, G. 2013. *Android Design Patterns: Interaction Design Solutions for Developers*. John Wiley & Sons Inc. (Indianapolis, Indiana, Mar, 2013).
- [10] Potel, M. 1996. *MVP: Model View Presenter The Taligent Programming Model for C++ and Java*. Taligent Inc.
- [11] van Krevelen D. W. F., Poelman R. A Survey of Augmented Reality Technologies, Applications and Limitations. *The International Journal of Virtual Reality*. Vol. 9, No. 2. (June 2010), 1-20.
- [12] Wasserman, Anthony I. 2010. Software Engineering Issues for Mobile Application Development. *Proceedings of the FSE/SDP Workshop on Future of Software Engineering Research* (Santa Fe, New Mexico, USA), 397-400.

Waldo: Serviço para Publicação e Descoberta de Produtores de Dados para *Middleware* de Cidades Inteligentes

Alternative Title: Waldo: Data Producers Registry and Discovery Service for Smart Cities Middleware

Marcelo Iury S. Oliveira
Centro de Informática, UFPE
UFRPE
Recife - PE, Brasil
miso@cin.ufpe.br

Kiev Santos da Gama
Centro de Informática, UFPE
Recife - PE, Brasil
kiev@cin.ufpe.br

Bernadette Farias Lóscio
Centro de Informática, UFPE
Recife - PE, Brasil
bfl@cin.ufpe.br

RESUMO

O conceito de Cidades Inteligentes tem despontado como a combinação do ambiente digital com comunidades reais, tornando mais eficiente a gestão dos espaços urbanos. A maioria das soluções propostas está voltada para a proposição de *middleware* de interconexão de dispositivos, coleta e mediação de dados, carecendo de mecanismos eficientes para a publicação e descoberta de produtores de dados. Este artigo apresenta o Waldo, um serviço de publicação e descoberta de produtores de dados voltado para o contexto de Cidades Inteligentes. O projeto do Waldo foi fundamentado no paradigma de Computação Orientada a Serviços, em soluções de bancos de dados NoSQL e na especificação SensorML. Através de simulações pode-se verificar a viabilidade do Waldo como solução para a publicação e descoberta de produtores heterogêneos.

Palavras-chave

Cidades inteligentes, serviço de descoberta, orientação a serviço, banco de dados NoSQL.

ABSTRACT

The smart cities concept has emerged from the combination of digital environments with real communities, thus making more efficient the urban spaces management. The majority of proposed solutions are focused on the proposition of middleware that interconnects devices, collects and mediates data. However, these solutions lack of effective mechanisms for publication and discovery of data producers. This paper presents the Waldo, a service for registry and discovery of data producers in the context of smart cities. Waldo follows the Service-Oriented Computing paradigm and adopts NoSQL database solutions and SensorML specification. Through simulations, we attested the feasibility of Waldo as a solution for publishing and discovery of heterogeneous data producers.

Categories and Subject Descriptors

H.2.4 [Database Management] - Systems - Parallel databases, Query processing, Distributed databases. H.4.m [Information Systems]: Miscellaneous

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.
Copyright SBC 2015.

General Terms

Algorithms, Management, Standardization, Design

Keywords

Smart Cities, discovery service, service oriented computing, NoSQL databases systems

1. INTRODUÇÃO

A evolução das tecnologias de informação e comunicação (TICs) permitiu o desenvolvimento de novas formas de geração, aquisição e processamento de dados. Seguindo esta tendência, cidades também estão fazendo uso das TICs para tornar mais eficiente a gestão dos seus espaços urbanos, com o intuito de melhorar a qualidade de vida de seus habitantes, assim como aperfeiçoar o uso dos equipamentos e recursos públicos, como também preservar o meio ambiente. Este novo modelo de construção e gestão dos espaços urbanos é chamado de Cidades Inteligentes (do inglês, *Smart Cities*) [1, 2].

Soluções de Cidades Inteligentes, além da automatização de subsistemas específicos, a exemplo das aplicações UbiBus [8], Busão [10], PowerMatcher [9], devem buscar estabelecer a integração de dispositivos e sistemas individuais de forma a viabilizar visões mais abrangentes e um contexto integrado da cidade. Assim, um projeto de cidade inteligente deve ser pensado como uma rede de inter-relacionamentos, um sistema interligado, mais orgânico, no qual os relacionamentos são tão importantes quanto às partes [1].

Os sistemas e dispositivos participantes de uma solução de Cidade Inteligente podem atuar tanto como produtores de dados (PDs) quanto consumidores de dados (CD). Os produtores de dados podem ser qualquer tipo de recurso, físico ou virtual, que seja capaz de prover dados para os consumidores de dados. Os dados, neste caso, podem advir de sensores, *smartphones*, até de bases de dados abertos e aplicações da Web.

Diversas soluções têm sido propostas para tratar a integração de produtores e consumidores de dados no contexto de Cidades Inteligentes. Uma parte significativa destas soluções está centrada no uso de um *middleware* de interconexão de dispositivos, coleta e mediação de dados. Este tipo de *middleware* tem o objetivo de viabilizar o desenvolvimento rápido de aplicações urbanas, lidando com diversos tipos de heterogeneidade, incluindo os diferentes níveis de heterogeneidade de dados (estrutural, sintática e semântica) [18].

Apesar de serem eficientes no que tange a integração de dispositivos e mediação dos dados [3], estas soluções carecem de mecanismos eficientes para a publicação e descoberta de produtores de dados. Um serviço de publicação e descoberta deve permitir que produtores sejam

capazes de publicar detalhes sobre seus serviços para que possam ser encontrados por consumidores de dados. E, alternadamente, o serviço de descoberta deve permitir que consumidores sejam capazes de realizar buscas e facilmente descobrir os produtores mais adequados às suas necessidades.. Uma vez descobertos os produtores, o serviço de descoberta deve retornar aos consumidores dados suficientes para que possam avaliar os produtores de forma independentemente, antes de contatá-los ou integrá-los às suas aplicações.

Como a quantidade de potenciais recursos que podem atuar como produtor de dados em uma cidade inteligente é massiva e crescente, selecionar qual o produtor mais apropriado para o provimento de determinados dados se torna um novo desafio a ser resolvido. No mais, o problema da publicação e descoberta requer também a resolução de questões, tais como recursos autônomos e heterogêneos, ausência de estruturação ou padronização e interoperabilidade [4].

Este artigo propõe um serviço de descoberta de produtores de dados, denominado Waldo. O serviço proposto faz uso de um conjunto de especificações abertas, vocabulários de domínio e de banco de dados NoSQL a fim de facilitar a publicação e descoberta de produtores de dados para qualquer *middleware* de Cidades Inteligentes. Este serviço auxiliará o processo de desenvolvimento de novas aplicações construídas a partir de dados urbanos, visto que facilitará a descoberta de produtores de dados que possam atuar como fontes de dados para consumidores de dados. A viabilidade da solução proposta foi validada através de uma prova de conceito que consistiu da implementação de um protótipo da solução e avaliação do protótipo por meio da simulação de um ambiente real inspirado na cidade do Rio de Janeiro.

A organização deste documento segue da seguinte forma: após esta seção introdutória, na qual é feita uma explanação geral sobre a problemática que motiva a realização deste trabalho. Na Seção 2, é apresentado o método de pesquisa. Na Seção 3, são apresentados os conceitos que caracterizam a solução proposta, assim como, sua arquitetura e projeto. A avaliação do mecanismo proposto é apresentada na Seção 4 e, na Seção 5, são apresentados alguns trabalhos relacionados. A Seção 6 conclui o trabalho, com um breve sumário das contribuições e apontando possíveis trabalhos futuros.

2. MÉTODO DE PESQUISA

O presente trabalho foi guiado pela postura filosófica positivista que definiu as estratégias de investigação e métodos de pesquisa. O pensamento positivista é orientado pela busca da construção do conhecimento a partir de inferências lógicas de um fato observado por meio de uma visão determinista, racional e cartesiana. A escolha do positivismo se deve ao fato da área de pesquisa Cidades Inteligentes ainda estar em estágios iniciais, fazendo com que o reducionismo permita uma análise mais apropriada de um contexto complexo que ainda carece de teorias e conceitos amplamente aceitos.

Sob o ponto de vista de seus objetivos, este trabalho tem caráter exploratório, na medida em que busca proporcionar maior entendimento sobre um problema [11]. Especificamente, este trabalho busca investigar, projetar e desenvolver um serviço que permita a publicação de produtores heterogêneos e autônomos, assim como a descoberta dos mesmos por parte de consumidores de dados também heterogêneos e autônomos. Diante do exposto, o presente trabalho possui a seguinte pergunta de pesquisa.

A abordagem de registro e descoberta de produtores de dados proposta para o Waldo é uma solução viável para o contexto de Cidades Inteligentes?

Segundo Yin, a escolha do método de pesquisa está relacionada com aspectos relativos ao tipo de investigação que se pretende executar [11]. Neste sentido, devem ser considerados fatores, tais como tipo de questão de pesquisa e o nível de controle sobre a unidade de pesquisa e suas variáveis [11]. A pergunta de pesquisa proposta possui caráter causal. Este tipo de pergunta, usualmente, é respondida por meio de experimentos. Contudo, no contexto de Cidades Inteligentes, a realização de experimentos usando sistemas e componentes reais é muito complexa de ser viabilizada por causa de fatores como dificuldade de acesso a sistemas reais em funcionamento, ou simplesmente ausência de tais sistemas.

Sendo assim, neste trabalho, foi realizada uma prova de conceito. Provas de conceitos consistem da implementação de protótipos com as funções fundamentais da solução almejada, seguida de uma avaliação a fim de analisar a viabilidade das tecnologias. Além disso, busca-se aferir a complexidade, tempo necessário para o desenvolvimento completo, os custos envolvidos, assim como métricas de qualidade ou desempenho de avaliação para a solução [27]. No caso deste trabalho, a avaliação do protótipo foi realizada por meio de simulação.

3. SOLUÇÃO PROPOSTA

Este trabalho propõe um serviço, intitulado Waldo, para a publicação e descoberta de produtores de dados. O Waldo é um serviço independente de plataformas tecnológicas, podendo ser conectado a qualquer *middleware* de Cidades Inteligentes. Ademais, considerando que o universo de potenciais produtores é massivo, dinâmico e heterogêneo, o Waldo também tem como objetivos não funcionais ser escalável para atender a demanda crescente de produtores de dados; ser robusto para lidar com a dinamicidade dos participantes e ser extensível para lidar com a heterogeneidade.

O projeto do Waldo foi fundamentado nos princípios da Computação Orientada a Serviços (COS), que é um paradigma que determina que as funcionalidades implementadas pelas aplicações devem ser disponibilizadas na forma de serviços [14]. Serviços são módulos que possuem interfaces formalmente definidas que permitem o acesso remoto de seus métodos e a transmissão de dados por meio de trocas de mensagens. As interfaces, assim, abstraem a implementação do serviço. Inclusive, a COS prevê a utilização de um mecanismo de registro e descoberta de serviços que resulta em um tripé bem definido de provedor de serviço, cliente e catálogo.

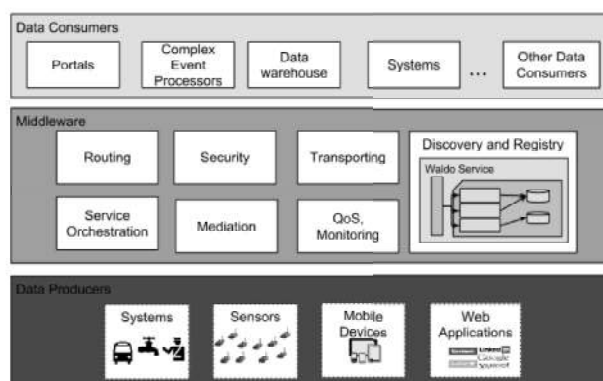


Figura 1. Integração do Waldo com *Middleware*

A abordagem COS traz duas vantagens adicionais. A primeira é permitir que o Waldo seja incorporado a qualquer *middleware* genérico de Cidades Inteligentes, tal qual o apresentado na Figura 1. A incorporação poderá ser realizada sem a necessidade de completa reimplementação de ambas as soluções, *middleware* e Waldo, podendo este último ser acoplado aos módulos de Registro e Descoberta do *middleware*. A segunda vantagem é que o Waldo também pode ser

agregado a outros catálogos de produtores de dados já existentes, permitindo, dessa forma, a troca de dados entre iniciativas de cidades inteligentes distintas, podendo atuar, assim, como um produtor de dados.

3.1 Arquitetura do Serviço Waldo

A arquitetura do serviço Waldo, ilustrada na Figura 2, possui dois catálogos para armazenamentos de metadados e vocabulários, três módulos principais (*i.e. Registry, Status Monitor e Vocabulary Registry*) e dois componentes auxiliares (*i.e. Adapter e Facade*).

Os catálogos são usados para armazenamento de informações sobre todos os PDs. O **PD Catalog** é usado para persistir os metadados dos PDs registrados no Waldo. Por sua vez, o **Vocabulary Catalog** é usado como dicionário de vocabulários usados para descrição dos PDs e de seus dados transmitidos pelos barramentos.

O **Registry** será responsável pelas operações de criação, atualização, remoção e consultas de instâncias de PD. Especificamente, as operações de consulta devem permitir a busca por PD de acordo com parâmetros, tais como fenômeno de interesse, limitação espacial, temporal, ID do sensor entre outros. Há também uma operação de *lookup*, que ao informar o ID (*i.e. URI*) de um PD, será retornado a sua descrição completa.

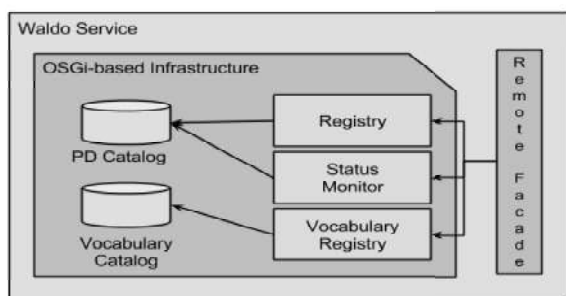


Figura 2. Arquitetura do Waldo

O **Registry** possui mecanismos de autorização para evitar que um participante não autorizado modifique ou remova um registro de PD. Um participante tem acesso apenas à porção de metadados para a qual possui permissão. O mecanismo de controle de acesso possibilita a manipulação dos metadados estritamente de acordo com a política de autorização especificada para o sistema. São pré-condições para o seu adequado funcionamento que o usuário que solicita acesso tenha sido corretamente autenticado.

O componente **Status Monitor** é o módulo responsável por monitorar a disponibilidade dos PDs. Como os PDs participantes de um barramento possuem comportamento autônomo, ocasionalmente, um PD pode se tornar temporariamente ou permanentemente indisponível. Caso o PD não solicite a remoção do registro, um CD pode, sem sucesso, tentar contatá-lo. Para evitar estes casos, o Status Monitor concede um "período de graça" durante o qual um registro de PD pode permanecer ativo no sistema. Quando expirado este período, o Status Monitor tenta contatar o PD. Caso a comunicação seja mal-sucedida, o Status Monitor marca o PD como *unavailable*. Se um PD marcado como *unavailable* voltar a se conectar no sistema, seu registro será marcado como disponível novamente.

O módulo **Vocabulary Registry** armazena vocabulários semânticos usados na descrição de PDs. A ideia é que usuários utilizem este módulo para descobrir as definições semânticas para os termos, conceitos e fenômenos tratados por um PD. Desta forma, o Vocabulary Registry auxilia tanto o processo de registro quanto de descoberta de PDs. Internamente, o Vocabulary Registry possuirá previamente

cadastrado um conjunto de vocabulários mais utilizados no contexto de Cidades Inteligentes. Contudo, novos vocabulários poderão ser cadastrados para suprir a ausência de definições semânticas.

Ainda, o Vocabulary Registry oferece basicamente quatro métodos: três de consulta, e um para registro. Os métodos de consultas simples permitirão encontrar a definição semântica de um termo com base em uma URI (*Uniform Resource Identifier*) ou em palavras-chaves. O método de registro permitirá o cadastro de novos vocabulários, podendo ser empregado algum método de moderação a fim de evitar inconsistências na base.

Os módulos auxiliares são na verdade implementações de padrões de projeto bem conhecidos e bastante utilizados em diversos sistemas. O componente **Facade** tem o objetivo de reunir as funcionalidades do Waldo, disponibilizando uma interface unificada para o conjunto de serviços providos pelos componentes Registry, Status Monitor e Vocabulary Registry. A **Facade** cria uma camada mais alta que facilita o uso dos componentes principais. A **Facade** poderá possuir um ou mais *proxies* remotos para permitir a invocação remota de suas funcionalidades por meio do modelo de comunicação *Request/Reply*. Desta forma, tanto o PD quanto o CD somente podem se relacionar com o Waldo através da **Facade** que, ao receber uma requisição, encaminha-as para o módulo responsável por atendê-las.

Por sua vez, o componente **Adapter** tem três funções. A primeira é permitir que PDs com interfaces incompatíveis consigam se conectar ao barramento e transmitir seus dados. No mais, os dados também poderão ser estruturados e processados pelo *adapter* para um formato mais apropriado para transmissão no barramento. A segunda função é permitir a agregação de um conjunto de dispositivos ou fontes de dados como um único PD, facilitando, assim, a integração, a exemplo de uma rede de sensores no barramento. A última função do **Adapter** é padronizar a forma de requisição de metadados e informações de um PD.

3.1.1 Relacionamento entre os Módulos

Os componentes principais e os auxiliares se relacionam como ilustrado na Figura 3. Todos os componentes, principais e auxiliares, são encapsulados como serviços. Basicamente, a interação se dá por intermédio de quatro relacionamentos, sendo eles:

1. Os PDs se registram no Waldo ao fazer uma submissão de requisição de cadastro na **Facade**, fornecendo informações relevantes quanto à utilização do mesmo, tais como *endpoint* de acesso, protocolos de transporte suportados, interfaces de interação, formato dos dados de entrada e saída. **Facade** encaminha a requisição para ser processada pelo Registry;
2. Os CDs solicitam ao Waldo a descoberta de PDs através da submissão de requisições de busca por PDs que satisfaçam certos parâmetros, tais como categoria, funcionalidade e características. A **Facade** ao receber a requisição encaminha para o Registry devolve uma lista com PDs que satisfazem às condições estabelecidas na requisição de descoberta;
3. O CD, após escolher um serviço dentre aqueles retornados pelo Registry, inicia o relacionamento com o PD provedor do serviço almejado. Requisições de dados serão enviadas ao respectivo PD por meio de sua interface **Adapter**. O retorno dos dados poderá ser encaminhado diretamente ao CD ou mediado por meio de uma interface de dados usada pelo middleware de cidade inteligente adjacente;
4. O Status Monitor ocasionalmente envia requisições de checagem para PDs que não atualizaram seus registros dentro do "período de graça" concedidos a eles.

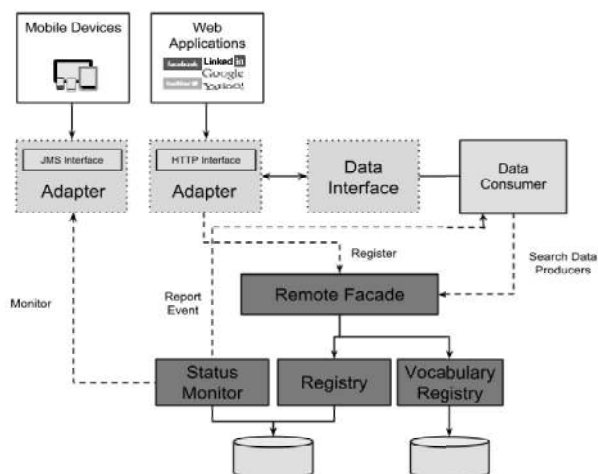


Figura 3. Interação entre os componentes

3.2 Escalabilidade e OSGi

Soluções de Cidades Inteligentes tem escalabilidade e robustez como requisitos inerentes a fim de suportar o aumento de carga devido ao crescimento da quantidade de produtores e consumidores de dados. O encapsulamento das funcionalidades do Waldo como serviços contribui para atender os requisitos de escalabilidade na medida em que aumenta a manutenibilidade. Contudo, essa abordagem não necessariamente permite atualização ou replicação de serviços de forma dinâmica. Assim, é necessário o uso de uma plataforma na qual serviços possam ser adicionados, atualizados e removidos em um ambiente de forma dinâmica em tempo real.

A solução proposta para o Waldo consiste na utilização de OSGi [16], uma plataforma que suporta tanto o desenvolvimento de aplicações a partir da composição de módulos reutilizáveis quanto o carregamento dinâmico de aplicações em tempo de execução. Os módulos possuem informações sobre os serviços que fornecem e sobre suas dependências que devem ser instanciadas ou importadas para o correto funcionamento.

Através do OSGi, novas instâncias (*i.e.* réplicas) dos módulos do Waldo podem ser disponibilizadas sem interrupções e em tempo de execução para lidar com a variação da carga. O provimento destas novas instâncias pode ser realizado pela alocação de novas máquinas sejam elas físicas ou virtuais. Ambas as abordagens permitem a chamada escalabilidade horizontal.

Desta forma, a *Facade* passa a incorporar uma fila de requisições e um conjunto de *threads* despachantes. Todas as requisições recebidas pela *Facade* são armazenadas em uma fila que é consumida pelas *threads* despachantes que, por sua vez, encaminharão as requisições para uma das instâncias disponíveis dos módulos principais. Por sua vez, este encaminhamento é também realizado pela invocação remota dos módulos.

3.3 Metadados de Descrição dos Produtores de Dados

No que tange a heterogeneidade, a arquitetura proposta deve ser flexível para permitir tanto o registro de produtores heterogêneos de dados, quanto o atendimento as requisições de busca oriundas de consumidores de dados também heterogêneos.

O Waldo faz uso do padrão *Sensor Web Enablement* (SWE) do grupo OGC (*Open Geographic Consortium*) [5] para descrever um PD, bem como o tipo de dado que produz. Apesar de haver outras estratégias

propostas para tratar da heterogeneidade e interoperabilidade entre dados de sensores e qualquer tipo de aplicação, o SWE é uma das estratégias mais robustas e aceitas pela comunidade científica [13]. Vale ressaltar que o SWE não é um sistema, mas sim uma arquitetura que define padrões para formatos de dados, metadados e interfaces de serviços.

Especificamente, o Waldo utiliza a especificação *Sensor Model Language* (SensorML) para modelar um PD. Além disso, a especificação *Observations & Measurements* (O&M) será usada como um dos vocabulários para definição dos atributos de descrição de um registro de PD, a exemplo da descrição de parâmetros, outputs, localização, identificação, entre outros. Os outros padrões do SWE (*e.g.* SOS, SAS) definem as interfaces de componentes essenciais para o gerenciamento de sensores que não dizem respeito ao cerne do Waldo, não sendo, desta forma, usados por ele.

O Waldo não faz uso da especificação completa de SensorML na modelagem de um PD. O SensorML define diversos elementos/entidades, tais como *PhysicalComponent*, *PhysicalSystem*, *ProcessChain*, *Detector*, *Sensor* e outros. Porém, apenas o elemento *PhysicalComponent* é usado pelo Waldo. Esta entidade foi escolhida, pois permite a descrição de atributos de identificação, caracterização, funcionalidades, outputs, parâmetros e serviços. Estes atributos foram considerados suficientes para permitir a descoberta de PDs. Apesar de o SensorML representar os dados por meio de XML, o Waldo permite o registro, atualização e busca de PDs recebendo como parâmetros tanto objetos Java quanto dados em XML ou JSON.

4. PROVA DE CONCEITO

Nesta seção, serão apresentados detalhes sobre o processo de avaliação da solução proposta, incluindo apresentação das tecnologias utilizadas para o desenvolvimento do protótipo e a descrição do ambiente de simulação utilizado na avaliação.

4.1 Protótipo e Tecnologias Utilizadas

Para validar a solução proposta, foi desenvolvido um protótipo em linguagem Java, tendo sido escolhida por apresentar facilidades para a implementação do sistema. Dentre as facilidades providas, podem-se destacar: alta portabilidade do código-fonte, existência de suporte para comunicação através da rede, bibliotecas nativas para acesso a banco de dados e suporte a programação concorrente.

No mais, optou-se por usar o *framework* Apache CxF [15] para o desenvolvimento dos *frontends* remotos do Waldo. O Apache CxF é um *middleware* de código aberto que ajuda no desenvolvimento de serviços usando APIs de programação remota. O Apache CxF facilita o processo de integração do Waldo com os *middlewares* existentes de Cidades Inteligentes na medida em que oferece suporte para diversos tipos de protocolos como SOAP, XML/HTTP, HTTP RESTful, ou CORBA.

Utilizou-se a solução de banco de dados NoSQL MongoDB para a implementação do PD Catalog. O MongoDB apresenta características próprias ao contexto de Cidades Inteligentes tais como alta disponibilidade, menor tempo de resposta, suporte a paralelismo e flexibilidade de esquema [19]. Em especial, o MongoDB oferece suporte nativo à replicação automática e distribuição horizontal, permitindo, assim, atender mais facilmente requisitos de escalabilidade. Além disso, a não obrigatoriedade de utilização de um esquema comum para todos os registros, permite tanto a extensão e atualização dos esquemas de representação dos PDs, quanto o suporte a registros esparsos nos quais alguns campos não são preenchidos.

Conjuntamente com o MongoDB, foi utilizado o *framework* Apache JCS [24] para implementação do sistema de *cache* de consultas. JCS é um sistema de *cache* distribuído escrito em Java. Foi especialmente

construído para acelerar aplicações, fornecendo um meio para gerenciar dados em *cache* de diversas naturezas dinâmicas.

Foi utilizada a biblioteca SWE *Common Library* [20] como base para a implementação das funções de codificação, decodificação e validação dos registros dos PDs descritos em SensorML. Em virtude de o SensorML ser especificado em XML e o MongoDB utilizar um formato derivado do JSON, foi necessária a implementação de *parsers* que permitissem a conversão entre os dois formatos. A propósito, empregou-se a convenção *Badgerfish* [21] na tradução de XML e JSON. Essa convenção estabelece um conjunto de regras que, sobretudo, lida com os aspectos de conversão de atributos, elementos vazios e a ordem de disposição das informações.

4.2 Avaliação

O processo de avaliação teve como principal objetivo responder a pergunta de pesquisa apresentada na Seção 2. Porém, é importante notar que a análise de viabilidade envolve a avaliação de um conjunto de parâmetros quantitativos e qualitativos, a exemplo de custo-benefício, manunibilidade e segurança. Embora existam diversos fatores, no contexto de Cidades Inteligentes, a heterogeneidade e o volume massivo de produtores e consumidores fazem com que robustez e flexibilidade sejam requisitos essenciais para o atendimento da viabilidade de uma solução de registro e descoberta.

A robustez diz respeito a capacidade de atendimento a grandes volumes de requisições, sejam elas de registro ou descoberta, com os devidos critérios de desempenho e disponibilidade. Por sua vez, flexibilidade diz respeito a capacidade de suportar a ausência de padronização entre os dispositivos, seja em relação as suas características, aos serviços ofertados, aos protocolos de comunicação ou aos modelos de dados usados. Desta forma, esses requisitos derivaram as seguintes perguntas de pesquisa secundárias para avaliação da viabilidade:

1. Robustez:
 - a) O Waldo é capaz de manipular requisições frequentes de cadastro e descoberta de PDs?
2. Flexibilidade
 - a) O Waldo suporta o registro de PDs heterogêneos?
 - b) O Waldo suporta a incorporação ou remoção de atributos de um PD?

Uma vez definidas as perguntas de pesquisa, é preciso projetar os experimentos que serão usados para respondê-las. Contudo, um problema que surge ao se projetar experimentos para soluções de Cidades Inteligentes é a complexidade em controlar e viabilizar as variáveis e unidades experimentais. A realização de experimentos usando sistemas e componentes reais pode ser impossível ou impraticável por causa do alto custo de prototipagem e testes, ou por causa da duração do experimento em tempo real ser impraticável.

A alternativa usada neste trabalho foi o emprego de simulação como um meio para a realização de experimentos. A simulação permite que se modele um sistema próximo do real, aumentando a confiabilidade nos resultados e nas decisões [5]. Desta forma, visando validar a solução proposta, foi desenvolvida uma aplicação que faz uso de *mocks* para simular a interação de PDs e CDs com o Waldo. Os *mocks* são objetos pré-programados que simulam partes do comportamento de um sistema real e são capazes de verificar se o sistema está funcionando conforme especificado.

Por fim, a avaliação do Waldo foi realizada utilizando-se o protótipo apresentado na Seção 4.1. As perguntas de pesquisa em estudo envolvem essencialmente a avaliação dos módulos *Facade*, *Registry*,

PD Catalog e *Adapter*. Os outros módulos, apesar de relevantes, não foram considerados na avaliação pelo fato de possuírem funcionalidades secundárias no Waldo.

4.2.1 Modelo de Simulação

O ambiente de simulação foi composto por um conjunto de *mocks* PDs e CDs que, de forma programada, submetiam requisições de cadastro, atualização e consulta para o Waldo. No que tange o lado servidor da solução, o simulador dispunha de um conjunto de *threads* responsáveis por tratar as requisições. Cada *thread* executava uma instância do Waldo.

A simulação era executada em ciclos, na qual um determinado conjunto de CDs ou PDs escalonava requisições que eram submetidas individualmente em um tempo aleatório máximo de 1 segundo. A quantidade de ciclos executados por cada conjunto era determinada pela quantidade de requisições a serem submetidas por cada entidade. Por exemplo, 10.000 consumidores enviando 100 consultas equivalem a 100 ciclos. Além disso, não havia sincronização entre os ciclos, assim podendo existir consultas oriundas de diferentes conjuntos sendo atendidas simultaneamente pelo Waldo.

Como o objetivo da simulação era verificar como o sistema se comportava frente ao recebimento intermitente de requisições, independentemente do tempo de transmissão das mesmas, foram abstraídos aspectos relativos à comunicação, tais como topologia de rede, congestionamento e variação de tempo de transmissão. Desta forma, todas as entidades participantes da simulação executaram em uma única máquina, estando plenamente disponíveis e podendo se comunicar entre si sem a ocorrência de falhas.

O resultado obtido na simulação é composto por dois valores: o tempo total de execução e a vazão de atendimento de consultas. O tempo total é medido após a instanciação do *testbed*, até o momento do atendimento de todos os ciclos de requisição.

4.2.2 Cenários Simulados

Nas simulações, foram usados tipos distintos de CDs e PDs. Estas entidades são comumente usadas na produção de sistemas para Cidades Inteligentes [1]. A quantidade de PDs, conforme apresentado no Quadro 1, foi determinada com base em estimativas relacionadas a cidade do Rio de Janeiro, tendo sido escolhida como referência tanto pela dimensão da cidade, que é segunda maior cidade do Brasil, quanto pela significativa disponibilidade de bases de dados abertas sobre o Rio de Janeiro¹.

Há um conjunto de CDs para cada tipo de PD usado no *testbed*. Isto é, existe um conjunto de CD que buscam apenas Táxis, outros que buscam apenas ônibus e assim por diante. Todos CDs submetem requisições de descoberta de PDs que possuam um determinado valor, entre 8 valores possíveis, para o atributo WALDO_TESTE. Este atributo pertence ao grupo de atributos de características presente em todos os PDs. Desta forma, as requisições de descoberta submetidas na simulação recebiam como parâmetros o tipo de entidade a ser buscada e um atributo de característica a ser usado como predicado de filtragem. No total, foram geradas 80 consultas distintas.

A quantidade de tipos distintos de CDs foi proporcional a quantidade total de PDs e CDs disponíveis. Por exemplo, considerando um total de 10 mil CDs e 100 mil PDs, dos quais 34.230 (34,23%) representam PDs de Táxis, houve 3.423 CDs efetuando buscas relativas a táxis. Isto é, dos 10 mil consumidores, 34,23% deles efetuaram buscas sobre táxis, seguindo a mesma lógica, 900 (9%) e 160 (1,6%) efetuaram de ônibus e câmeras de tráfego, respectivamente.

¹ <http://data.rio.rj.gov.br/>

As quantidades totais de consumidores utilizadas foram 10.000 e 25.000. Estes valores foram definidos com base na quantidade média de usuários ativos do Waze em grandes cidades da Europa [24]. No mais, as simulações ocorriam com um percentual dos PDs já cadastrados e, ao longo da simulação, os PDs não registrados solicitavam o cadastro de seus metadados.

Quadro 1. Sumário de Produtores de Dados Usados na Avaliação

Tipo de Produtores	Atributos do Produtor	Quantidade de PDs
Táxi	Identificação, Localização, Características e Output	34.230
Ônibus	Identificação, Localização, Características e Output	9.000
Estação Meteorológica	Identificação, Localização, Características, Funcionalidades e Output	320
Câmera de Tráfego	Identificação, Localização, Características, Funcionalidades, Input e Output	1.600
Base de Dados Aberta	Identificação, Características, Funcionalidades, Informações de Contato e Output	28.500
Semáforos de Trânsito	Identificação, Localização, Características, Funcionalidades e Output	5.250
Viaturas de (Polícia e Trânsito)	Identificação, Localização, Características e Output	2.000
Metrô e Trem	Identificação, Localização, Características e Output	100
Sistemas de Alerta	Identificação, Localização, Características, Funcionalidades e Output	320
Outras Fontes (páginas web e aplicativos sociais)	Identificação, Características, Funcionalidades, Informações de Contato e Output	18.680
Total		100.000

De maneira a avaliar a flexibilidade da solução, utilizaram-se esquemas distintos para representação dos PDs que compunham o *testbed*, como pode ser visto na coluna Atributos do Quadro 1. Todos os PDs possuíam em comum a descrição de sua identificação, características e output, podendo apresentar outras descrições tais como localização e capacidades.

Quadro 2. Cenários Simulados

Id. Cenário	Quant. de CDs	Quant. de Ciclos	Política de BD
Cenário 1	10 mil	100	Sem índices
Cenário 2	10 mil	100	Com índices
Cenário 3	10 mil	500	Com índices
Cenário 4	10 mil	100	Com índices, subcoleções
Cenário 5	10 mil	500	Com índices, subcoleções
Cenário 6	10 mil	100	Com índices, subcoleções e <i>cache</i>
Cenário 6	10 mil	500	Com índices, subcoleções e <i>cache</i>
Cenário 7	25 mil	100	Com índices, subcoleções e <i>cache</i>
Cenário 8	25 mil	500	Com índices, subcoleções e <i>cache</i>

Visto que a solução de banco de dados tem um impacto significativo para o desempenho do Waldo, foram avaliadas diferentes configurações do MongoDB com o objetivo definir os limites do sistema de acordo com a máquina disponível e como o sistema responde a carga de trabalho. As abordagens utilizadas foram: criação de índices, uso de subcoleções e de resultados.

Em todos os cenários executados, foram considerados um percentual de 50% dos PDs previamente cadastrados e uma limitação de 100 registros retornados por consultas. Esta limitação é uma prática comum em diversos sistemas da Web [23].

Além dos parâmetros já citados, que definem uma configuração base para a nossa avaliação, definimos diferentes cenários por meio da variação de dois outros parâmetros do modelo de simulação: quantidade de CDs e quantidade de ciclos de requisições. O Quadro 2 apresenta os cenários executados.

Todos os experimentos foram executados em um notebook Asus S46C, - Processador Intel Core i5-3317U 1.7GHz, Memória RAM 6GB e Sistema Operacional Windows 8.1 64 bits. Cada experimento executou um mínimo de 4 rodadas para cada cenário executado. Para grande maioria dos cenários, as 4 rodadas foram suficientes para obtermos resultados com 95% de nível de confiança e erro máximo de 5%.

4.2.3 Resultados

Durante a realização dos experimentos, o Waldo executou uma carga de trabalho entre 1 milhão a aproximadamente 12 milhões de consultas. Em relação ao consumo de memória, o Waldo manteve um consumo de memória estável em torno de 2 GB. A memória RAM é um dos principais pontos de análise em um mecanismo de descoberta, pois está diretamente ligada ao número de requisições que o servidor é capaz de atender em determinado tempo. Além disso, todos os 100 mil PDs do *testbed* ocuparam apenas 300MB de dados, possuindo uma média de 4 KB por registro. Em virtude de tratar-se também de um sistema de registro de dados, o baixo consumo do Waldo para o armazenamento dos registros torna viável manipular uma quantidade maior de produtores de dados pertencentes a uma cidade inteligente.

No mais, apesar de consumir toda capacidade de processamento da máquina de teste, a aplicação não apresentou problemas de concorrência durante os experimentos. Contudo, a duração e a vazão dos experimentos variaram de acordo com os cenários simulados. As Figuras 4 e 5 mostram os resultados obtidos com a simulação.

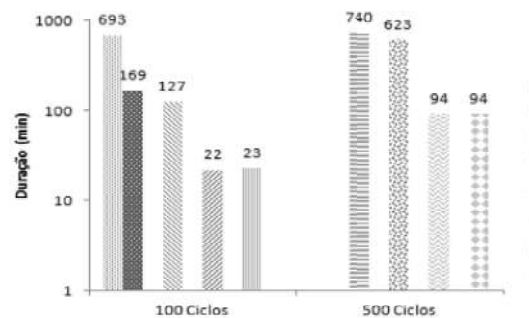


Figura 4. Duração da Execução

O Eixo y apresenta as latências e as vazões obtidas. Um resultado esperado que pode ser observado na Figura 4 é que, para a maioria dos cenários, a duração para o atendimento de todas as requisições aumenta à medida que as quantidades de consumidores e de ciclos de consultas também cresce. Entretanto, esse aumento da duração do experimento não está atrelado a uma melhoria na vazão do sistema. Pode-se perceber isso nos cenários 2 e 3, nos quais há apenas uma diferença de aproximadamente 9% (10 requisições), mesmo existindo aumento de 100 para 500 ciclos. Esse mesmo padrão pode ser percebido nos cenários 4 e 5.

Outro resultado que se destacou foi o ganho de eficiência com uso de *cache* para armazenamento das consultas, que apresentou resultados significativamente superiores ao restante das abordagens. Pode-se perceber isso fazendo uma análise dos cenários 5 ao 8. Nestes cenários,

o tempo máximo de duração foi de 23 e 94 minutos para, respectivamente, 100 e 500 ciclos. A mesma tendência pode ser observada na Figura 5. A menor vazão obtida com o uso de índices foi de 803 requisições/segundo. Esse resultado é aproximadamente 481% superior que a vazão obtida no cenário 4, que é o melhor dentre os quais não há uso de *cache*.

Analizando apenas os cenários nos quais há uso de *cache*, destaca-se outro resultado. Os cenários 7 e 8, apesar de possuírem uma carga de trabalho maior (25 mil consumidores em contraposição aos 10 mil consumidores utilizados nos outros cenários), apresentaram os melhores resultados em termos de duração e vazão, sendo de 1873 e 2223 requisições atendidas por segundo para 100 e 500 ciclos, respectivamente. Este resultado é explicado por dois fatores: pouca variação nas consultas submetidas e maior quantidade de consultas submetidas. O aumento do número de consumidores implica em mais consultas submetidas simultaneamente. Como o conjunto de tipos distintos de consultas é pequeno (*i.e.* 80 consultas), há um aumento da quantidade de *cache hit*, reduzindo a necessidade de busca em disco rígido e consequentemente diminuindo o tempo médio para atendimento das consultas.

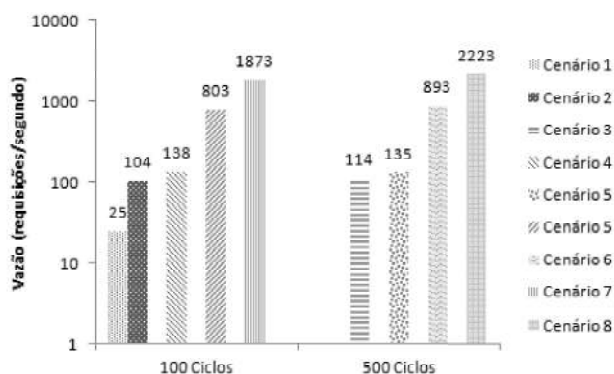


Figura 5. Vazão de Atendimento de Requisições

Aliado ao *cache*, a criação de índices se mostrou muito importante para alcançar um bom desempenho. Isso pode ser mais facilmente percebido no Cenário 1, que não fez uso de índices. Este cenário apresentou tempo de duração similar aos dos cenários 3, 5 e 7 que utilizaram índices e possuíam carga de trabalho 5 vezes superior ao Cenário 1. Por sinal, o uso de índices se torna mais determinante na medida em que a quantidade registros armazenados pelo Waldo cresce. Segundo os comandos de *profile* do MongoDB, as consultas executadas sem o suporte dos índices efetuam uma leitura completa dos registros existentes no sistema.

No que tange a avaliação da eficiência no tratamento de requisições de cadastro, a solução também apresentou bom desempenho. Durante a fase de preparação do *testbed*, que efetuava o registro de 50 mil PDs, também foi medida a vazão para o atendimento. O sistema apresentou uma vazão média de aproximadamente 4.800 requisições atendidas por segundo, tendo sido usadas 50 threads.

Em relação à avaliação da flexibilidade, o Waldo conseguiu lidar com o registro e busca de PDs heterogêneos. O SensorML, por ser baseado em XML e por oferecer um esquema plenamente extensível, permitiu a variação do conjunto de atributos de descrição, o preenchimento parcial de metadados, bem como, a adição e remoção de atributos. O MongoDB também contribuiu para o atendimento do requisito de flexibilidade ao suportar o registro de documentos heterogêneos (*i.e.* documentos com esquemas distintos) em uma mesma coleção.

No mais, em relação ao desempenho, não houve significativa diferença entre os tipos distintos de PDs registrados no sistema, embora

possuíssem esquemas de dados também distintos. Contudo, é preciso ressaltar que a estrutura de modelagem complexa do SensorML, tornou também complexa a criação do índices para otimização das consultas.

5. TRABALHOS RELACIONADOS

Algumas implementações do SWE foram propostas, dentre as mais robustas destacam-se os projetos OSIRIS² e GENESIS³. No entanto, estes projetos são restritos ao domínio de dados geoespaciais e a persistência é implementada com SGBD relacional. Dessa forma, além de não ser facilmente generalizável para outros tipos de aplicação, ainda há o problema de escalabilidade e rigidez no esquema de dados.

Um dos trabalhos mais recentes fundamentado no SWE foi desenvolvido por Fazio e colegas [6] e teve como objetivo propor uma arquitetura que permita o acesso a qualquer tipo de dados, obtido a partir de diferentes sites de sensoramento. Esta arquitetura utiliza um conjunto de linguagens baseadas em XML e especificações de interface de serviços Web (a exemplo, de REST) com o intuito de facilitar a descoberta, acesso e pesquisa sobre os dados dos sensores. Entretanto, a solução apresentada nesse trabalho parte do princípio de que os sites administrativos dos sensores serão responsáveis pela publicação e configuração dos dados. Além disso, o trabalho assume que os problemas de integração serão resolvidos na medida em que a estrutura e semântica dos dados esteja construída.

Tendo como ponto de partida a especificação do SWE, o W3C (*World Wide Web Consortium*) criou um grupo chamado *Semantic Sensor Networks Incubator Group* (SSN-XG) cuja tarefa foi desenvolver o *Semantic Sensor Network* (SSN) [7], que é uma ontologia criada para modelar sensores, sistemas, processos e observações. Enquanto os padrões especificados pelo SWE apenas proveem descrição para acesso a sensores e seus metadados, eles não oferecem a capacidade de abstração, categorização e raciocínio por meio de tecnologias semânticas [7].

No contexto de COS, há também diversos trabalhos que apresentam soluções de publicação e descoberta de serviços [25], dentre eles destaca-se o WS-Discovery [26] que define protocolos de descoberta através de trocas de mensagem baseadas em IP-Multicast. Apesar de ser bastante usado e de existirem diversas implementações consolidadas, o WS-Discovery não oferece nenhuma especificação ou *guideline* de como descrever um produtor de dados.

Por fim, dentre as propostas voltadas para sensores e internet das coisas, o EPCglobal, que é um conjunto de tecnologias que permite a identificação e o compartilhamento de informações a partir de *tags* RFID também possui um mecanismo de descoberta [28]. Contudo, o mesmo possui restrições de privacidade e alcance estando mais voltado para aplicações do tipo *Machine-to-Machine* [28].

6. CONCLUSÃO E TRABALHOS FUTUROS

Cidades inteligentes são iniciativas que buscam tornar a gestão de seus espaços e recursos mais eficientes por meio do uso de tecnologias de informação e comunicação. Diversas soluções tem sido propostas para viabilizar o desenvolvimento de Cidades Inteligentes com o intuito de estabelecer a integração de dispositivos e sistemas individuais diversos por meio do uso *middleware* de interconexão de dispositivos, coleta e mediação de dados. Contudo, essas soluções carecem de mecanismos eficientes para a publicação e descoberta de produtores de dados.

Este artigo apresentou o Waldo, um serviço de publicação e descoberta de produtores de dados voltado para o contexto de Cidades

² <http://www.osiris-fp6.eu>

³ <http://www.genesis-project.eu/>

Inteligentes. O Waldo possui um conjunto de componentes que permite aos consumidores de dados efetuarem a busca de produtores por meio de serviços simples, como busca por identificadores até a busca por produtores que atendam a determinados critérios, tais como tipo de dados fornecido e localidade do dispositivo. Para a avaliação do Waldo foi desenvolvido um protótipo em Java, auxiliado pelos *frameworks* Apache CxF e Apache JCS e por uma solução de banco de dados MongoDB.

As simulações confirmaram que o Waldo é uma solução viável quando usado o sistema de *cache* para armazenamento das consultas. Estes resultados também corroboram as suposições quanto à robustez e flexibilidade da solução no suporte a produtores e consumidores de dados heterogêneos.

É importante ressaltar que os resultados de desempenho apresentados neste trabalho também dependem diretamente da máquina utilizada. Os resultados apresentados neste artigo foram coletados através da execução do sistema em uma máquina doméstica. Certamente, em um servidor ou em um *cluster*, a capacidade do Waldo de atender consultas pode ser melhorada.

Dentre os trabalhos futuros, pode-se incluir a implementação do mecanismo de raciocínio do Vocabulary Registry que permitir localizar definições semânticas com base em palavras-chave ou definições semelhantes a uma definição usada como parâmetro. Outro desafio é fazer a extensão do SensorML para a incorporação de metadados relativo a contexto [29] e outros aspectos mais voltados para Cidades Inteligentes. Além disso, pretende-se, o estudo de outras soluções de banco de dados NoSQL para o armazenamento dos metadados, assim como a avaliação comparativa do Waldo em relação a outras soluções de publicação e descoberta de produtores de dados.

7. AGRADECIMENTOS

Este trabalho foi parcialmente suportado pelo Instituto Nacional de Ciência e Tecnologia para Engenharia de Software (INES). Marcelo Iury S. Oliveira é bolsista do CNPq-Brasil na modalidade doutorado.

8. REFERENCIAS

- [1] KOMNINOS, Nicos. Intelligent cities: innovation, knowledge systems, and digital spaces. Taylor & Francis, 2002.
- [2] SCHAFFERS, Hans et al. Smart cities and the future internet: Towards cooperation frameworks for open innovation. Springer Berlin Heidelberg, 2011.
- [3] GAMA, K., TOUSEAU, L., e DONSEZ, D. (2012). Combining heterogeneous service technologies for building an Internet of Things middleware. *Computer Communications*, 35(4), 405-417.
- [4] NAPHADE, Milind et al. Smarter cities and their innovation challenges. *Computer*, v. 44, n. 6, p. 32-39, 2011.
- [5] BOTTS, Mike et al. OGC® sensor web enablement: Overview and high level architecture. In: *GeoSensor networks*. Springer Berlin Heidelberg, 2008. p. 175-190.
- [6] FAZIO, Maria et al. Heterogeneous sensors become homogeneous things in smart cities. In: *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, 2012 Sixth International Conference on. IEEE, 2012. p. 775-780.
- [7] LEFORT, Laurent et al. Semantic sensor network xg final report. W3C Incubator Group Report, v. 28, 2011.
- [8] VIEIRA, Vaninha et al. The UbiBus project: Using context and ubiquitous computing to build advanced public transportation systems to support bus passengers. In: SBSI, 2012, São Paulo.
- [9] KOK, J. Koen; WARMER, Cor J.; KAMPHUIS, I. G. PowerMatcher: multiagent control in the electricity infrastructure. In: *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. ACM, 2005. p. 75-82.
- [10] LEITE, Daniel Farias Batista; ROCHA, Júlio Henrique; DE SOUZA BAPTISTA, Cláudio. Busão: um Sistema de Informações Móvel para Auxílio à Mobilidade Urbana Através do Uso de Transporte Coletivo. In: SBSI, 2013, João Pessoa.
- [11] YIN, Robert K. Estudo de Caso: Planejamento e Métodos. Bookman editora, 2015.
- [12] KAMIENSKI, Carlos et al. Colaboração na internet e a tecnologia peer-to-peer. In: XXV Congresso da Sociedade Brasileira de Computação-SBC2005. 2005.
- [13] BRÖRING, Arne et al. New generation sensor web enablement. *Sensors*, v. 11, n. 3, p. 2652-2699, 2011.
- [14] PAPAOGLOU, Mike P. Service-oriented computing: Concepts, characteristics and directions. In: *Web Information Systems Engineering*, 2003. WISE 2003. *Proceedings of the Fourth International Conference on*. IEEE, 2003. p. 3-12.
- [15] APACHE, C. X. F. (2009). An open source service framework. See: <http://cxf.apache.org>, 111.
- [16] ALLIANCE, T. O. "OSGi Service Platform Core Specification, Version 4.3." (2011).
- [17] MongoDB. disponível em <http://www.mongodb.org>, acessado em 04-03-2015
- [18] HASLHOFER, Bernhard; KLAS, Wolfgang. A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys (CSUR)*, v. 42, n. 2, p. 7, 2010.
- [19] VIEIRA, Marcos Rodrigues et al. Bancos de Dados NoSQL: conceitos, ferramentas, linguagens e estudos de casos no contexto de Big Data. SBB. São Paulo, 2012.
- [20] Lib-swe-common, disponível em <https://github.com/sensiasoft/lib-sensorml>, acessado em 04-03-2015
- [21] BADGERFISH. disponível em <http://badgerfish.ning.com/> acessado em 04-03-2015
- [22] WAZE STATS. disponível em <http://wazestats.com/> acessado em 04-03-2015
- [23] VESDAPUNT, Norases; GARCIA-MOLINA, Hector. Identifying Users in Social Networks with Limited Information. 2014.
- [24] Apache JCF. disponível em <http://commons.apache.org/proper/commons-jcs/> acessado em 04-03-2015
- [25] MUKHOPADHYAY, Debajyoti; CHOUGULE, Archana. A survey on web service discovery approaches. In: *Advances in Computer Science, Engineering & Applications*. Springer Berlin Heidelberg, 2012. p. 1001-1012.
- [26] BEATTY, John. Web services dynamic discovery (ws-discovery). <http://msdn.microsoft.com/library/en-us/dnglobspec/html/ws-discovery1004.pdf>, 2004.
- [27] CHURCHMAN, Charles West; SCHAINBLATT, A. H. The researcher and the manager: A dialectic of implementation. *Management Science*, v. 11, n. 4, p. B-69-B-87, 1965.
- [28] FABIAN, Benjamin; GÜNTHER, Oliver. Security challenges of the EPCglobal network. *Communications of the ACM*, v. 52, n. 7, p. 121-125, 2009.
- [29] GU, Tao; PUNG, Hung Keng; ZHANG, Da Qing. A middleware for building context-aware mobile services. In: *Vehicular Technology Conference, 2004. VTC 2004-Spring*. 2004 IEEE 59th. IEEE, 2004. p. 2656-2660.

Arquitetura de Software de Referência para Sistemas de Informação Governamentais

Alternative title: Reference Software Architecture for Government Information Systems

Maurício Serrano

Universidade de Brasília – FGA
Área Especial de Indústria Proj. A
72.444-240 Gama/DF – Leste
+55 (61) 3107-8901
serrano@unb.br

Milene Serrano

Universidade de Brasília – FGA
Área Especial de Indústria Proj. A
72.444-240 Gama/DF – Leste
+55 (61) 3107-8901
mileneserrano@unb.br

André Cruz Cavalcante

Universidade de Brasília – FGA
Área Especial de Indústria Proj. A
72.444-240 Gama/DF – Leste
+55 (61) 3107-8901
andrelink14@gmail.com

RESUMO

Nesse artigo, é descrita uma Arquitetura de Software de Referência para um Órgão Público Brasileiro. Essa arquitetura está baseada no Paradigma Orientado à Convenção sobre Configuração, o qual tenta diminuir a quantidade de decisões a serem tomadas pelos desenvolvedores, ganhando simplicidade, sem perder flexibilidade e produtividade. A arquitetura lida ainda com auditoria, rastreabilidade, autenticação, monitoramento e outros tópicos relevantes para sistemas de informação de Órgãos Públicos Brasileiros. Por fim, a arquitetura tem sido aplicada no desenvolvimento de dois sistemas de informação associados a um Órgão Público Brasileiro. Um desses é, atualmente, um dos mais importantes sistemas de informação do Órgão em questão.

Palavras-Chave

Arquitetura de Software de Referência, Reutilização, Governo, Padrão, Auditoria, Rastreabilidade, Autenticação, Monitoramento.

ABSTRACT

In this paper, we describe a Reference Software Architecture for a Brazilian Public Entity based on a software design paradigm, called Convention over Configuration, which tries to decrease the decisions made by developers, gaining simplicity, but not necessarily losing flexibility or productivity. Moreover, the proposed architecture deals with *auditing, tracing, logging, monitoring* and other relevant issues for information systems in the Brazilian government scope. Our Reference Software Architecture has been applying into two information systems' development at a Brazilian Public Entity. One of them is the most important information system at this Entity, nowadays.

Categories and Subject Descriptors

D.2 [SOFTWARE ENGINEERING]: D.2.1 [Requirements/Specifications] – *Elicitation methods; Methodologies; Tools*; D.2.11 [Software Architectures] – *Data abstraction; Domain-specific architectures; Information hiding; Patterns*; D.2.13 [Reusable Software] – *Reusable libraries*.

General Terms

Design, Security, Standardization, Languages.

Keywords

Reference Software Architecture, Reuse, Government, Pattern, Auditing, Tracing, Logging, Monitoring.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, no post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26-29, 2015, Goiânia, Goiás, Brazil.

Copyright SBC 2015

1. INTRODUÇÃO

As autarquias governamentais são organizações complexas, cujos impactos de suas atividades possuem alcance amplo. Nesse contexto, esses Órgãos registram, consultam e manipulam informações de alta relevância, tanto para conferir ao cidadão a prestação de serviço público quanto para apoiar o Estado na tomada de decisões estratégicas. Diante de problemas decorrentes da falta de disponibilidade, integridade, confidencialidade, autenticidade e manutenibilidade nos sistemas de informação associados aos Órgãos governamentais, ressalta-se a importância de padronização na Arquitetura de Software. No entanto, têm-se poucos estudos nesse contexto. Tornam-se necessários, ainda, esforços na área da Engenharia de Software e afins acordando soluções específicas para essas organizações. Por exemplo, a eficácia do serviço público demanda intensa cooperação entre organizações, visando, dentre outras necessidades, a troca de informações. Adicionalmente, normativas e padrões específicos, estabelecidos no setor público brasileiro (foco desse trabalho) devem ser rigorosamente respeitados. Apenas para ilustrar: Órgãos e entidades integrantes do Sistema de Administração dos Recursos de Tecnologia da Informação (SISP) devem observar o ePING [1] - *Padrões de Interoperabilidade de Governo Eletrônico* - no planejamento da contratação, aquisição e atualização de sistemas de informação. O ePING define um conjunto mínimo de premissas, políticas e especificações técnicas, que regulamentam a utilização das tecnologias de informação no âmbito do Governo Federal Brasileiro e estabelecem condições de interação entre –sistemas de informação – com as demais esferas governamentais.

Diante do exposto, o grupo dos autores tem desempenhado esforços na especificação de uma Arquitetura de Software de Referência (ASR) para novas contratações de desenvolvimento de Sistemas de Informação Governamentais por terceirizadas. Esse grupo atua como uma frente de Arquitetura de Software em um projeto de pesquisa e desenvolvimento, uma parceria entre a Universidade de Brasília – FGA e um Órgão Público Brasileiro. Dentre as contribuições relevantes do grupo, tem-se, principalmente: a análise de portfólio de sistemas de informação do Órgão; a investigação constante em busca de soluções de projeto, implementação, teste e manutenção específicas às necessidades desse tipo de Órgão governamental; a versão 1.0 da especificação da ASR; a capacitação de terceirizadas para o desenvolvimento de sistemas de informação segundo essa arquitetura; e o apoio no desenvolvimento, por parte da terceirizada, de *releases* de dois novos sistemas. Um desses sistemas é de uso interno do Ministério e o outro é de grande porte, âmbito nacional, e considerado um dos mais importantes sistemas de informação associados ao Ministério em questão.

Esse artigo, organizado em seções, procura apresentar em linhas gerais: o processo de análise institucional junto ao Órgão (Seção 2), no intuito de estudar o portfólio de sistemas de informações e as arquiteturas associadas a esses sistemas; a apresentação da ASR para Sistemas de Informação Governamentais (Seção 3); o desenvolvimento de sistemas de informação governamentais orientados pela arquitetura bem como outras contribuições associadas (Seção 4); e, finalmente, as considerações finais e possibilidades de trabalhos futuros (Seção 5).

2. ANÁLISE DE PORTFÓLIO

Ao longo do processo de análise do portfólio de sistemas de informação do Ministério, foram identificados mais de cinquenta sistemas em produção. Desses sistemas em produção, 58% foram desenvolvidos usando como base uma ASR em Java JSF e Struts [2]; 28% com base em uma ASR em PHP e 14% com base em uma ASR em ASP. Portanto, predominantemente, os sistemas de informação em produção nessa Autarquia são em Java.

Adicionalmente, ainda procurando acordar conformidades e não conformidades com base no portfólio de sistemas de informação, foram realizadas várias reuniões e entrevistas bem como aplicadas técnicas de introspecção e observação junto aos membros do Órgão Público Brasileiro. Dentre os membros participantes nesse processo, têm-se: membros da TI do Ministério; e membros das terceirizadas (i.e. desenvolvimento, qualidade e infraestrutura).

Tomando como base o rastro dos relatos acordados nas reuniões e entrevistas, foi possível a identificação de várias não conformidades, tipicamente encontradas em Órgãos governamentais, tais como: falta de qualidade dos sistemas de informação desenvolvidos; baixa produtividade da Fábrica de Software; insatisfação constante dos clientes finais dos sistemas de informação; falta de controle em termos de auditabilidade, *tracing* e *logging*; não possibilidade de monitoramento nos sistemas de informação em vigência; não possibilidade de extensibilidade dos sistemas de informação desenvolvidos, não sendo possível, por exemplo, acessá-los usando dispositivos móveis; ausência de flexibilidade desses sistemas visando à integração bem como a interconexão entre sistemas; excesso de documentação (maioria delas incompleta e não útil sob o ponto de vista do cliente e/ou de manutenção); não possibilidade de avaliação plena do sistema no ato da entrega do mesmo por parte da Fábrica de Software, e dificuldade de implantação de metodologias ágeis no desenvolvimento dos sistemas.

Apesar das não conformidades mencionadas, os membros da TI do Ministério solicitaram o não rompimento com as tecnologias já estabelecidas no Órgão. A preocupação sempre foi evitar mudanças bruscas em termos dos editais vigentes com as terceirizadas e/ou processos longos de capacitação da equipe de infraestrutura e da Fábrica de Software.

O Coordenador Geral da TI do Ministério (CGTI), também participante no processo de análise, enfatizou três critérios de qualidade em particular, os quais deveriam ser contemplados na especificação da nova ASR: produtividade, flexibilidade e extensibilidade. As preocupações dos demais membros da TI do Órgão concentravam-se nas questões de auditabilidade, *tracing*, *logging* e monitoramento. Os membros de infraestrutura do Órgão sensibilizaram-se com a manutenibilidade evolutiva dos sistemas de informação bem como com questões de integração entre sistemas. Essa integração necessariamente deveria se orientar pelo ePING - *Padrões de Interoperabilidade de Governo Eletrônico*.

Conforme detalhado na próxima seção, a frente de Arquitetura de Software do projeto dedicou-se à especificação de uma nova ASR capaz de atender às expectativas dos interessados, considerando os pontos de vista analisados e um suporte tecnológico emergente.

3. ARQUITETURA DE SOFTWARE

3.1 Convenção sobre Configuração

O Paradigma Orientado à Convenção sobre Configuração [3] é um modelo de desenvolvimento emergente, que procura minimizar o número de decisões tomadas pelo time de desenvolvimento, simplificando o processo. Essa simplificação é provida devido ao uso de padrões e configurações. Essencialmente, estar orientado à convenção sobre configuração significa que o time de desenvolvimento precisa definir apenas aspectos não convencionais. Por exemplo, quais convenções adotar visando padronizar o código em termos de anotações.

Dessa forma, o desenvolvimento orientado à convenção sobre configuração, quando apoiado por uma ferramenta e/ou plataforma que colabora na definição do comportamento desenhado pelo time de desenvolvimento, tende a diminuir o esforço na redação de arquivos de configuração; facilitar aspectos de padronização; melhorar a produtividade do time, e proporcionar uma manutenibilidade facilitada.

3.2 Suporte Tecnológico

Atualmente, existem várias propostas de plataformas e *frameworks* que se orientam pelo Paradigma Orientado à Convenção sobre Configuração, provendo suporte à redação de arquivos de configuração. Nesse cenário, o grupo dos autores pesquisou vários suportes tecnológicos, com destaque para: Grails [4], RubyOnRails [5] e Play [6]. Todas são plataformas que conferem ao time de desenvolvimento um ambiente completo, orientado à convenção sobre configuração, com *plugins*, arquivos de configuração, geração automática de código (em alguns casos) e outros facilitadores que permitem aumentar a produtividade do time de desenvolvimento, mantendo a qualidade do software.

Os estudos, conduzidos com experimentações, provas de conceito, pesquisa-ação, levantamento bibliográfico, e outros suportes à tomada de decisão, permitiram à equipe optar pela plataforma de desenvolvimento Grails. Dentre as razões que justificaram essa escolha, têm-se:

- (i) o não rompimento com as tecnologias já utilizadas nas terceirizadas, prestadoras de serviços de desenvolvimento, qualidade e infraestrutura do Órgão Público em estudo. Esse tópico era uma grande preocupação dos membros da TI do Órgão, pois demandaria mudanças acentuadas na forma de trabalhar da Fábrica de Software; na geração dos relatórios de qualidade aferidos pela contratada de qualidade; nos ambientes de desenvolvimento, homologação, teste e produção mantidos pela equipe de infraestrutura associada ao Órgão, e nos editais de contratação vigentes na Autarquia. Portanto, optar por RubyOnRails seria adotar uma nova linguagem de programação, o Ruby. Essa linguagem não se assemelha à linguagem Java, predominantemente utilizada nos sistemas do Órgão. A adoção dessa plataforma demandaria maior esforço na capacitação das equipes, além de vários ajustes nos editais, infraestrutura e outros setores críticos da Autarquia. O Play faz uso da linguagem de programação Scala. Trata-se de uma linguagem objeto-funcional próxima à linguagem Java, inclusive compilada para Java bytecode e executável em uma

Máquina Virtual Java. O Grails usa a linguagem de programação Groovy, também compilada para Java bytecode e executável em uma Máquina Virtual Java. O Grails tem vantagens em relação ao Play por contar, principalmente, com toda a comunidade Java, a qual apoiou o desenvolvimento de *plugins* e outros suportes associados ao Grails. Adicionalmente, nossas pesquisas de tendência de mercado e na literatura [7] indicaram maior adesão da comunidade (academia e empresas) à plataforma Grails em detrimento do Play.

(ii) a necessidade de manter aspectos de auditabilidade, criação de perfis, *logging*, *tracing* e monitoramento para maior controle por parte dos clientes do sistema e pelos membros da TI do Órgão. Plataformas de desenvolvimento orientadas à convenção sobre configuração fazem uso de suportes reutilizáveis e de fácil configuração visando contemplar as aplicações de funcionalidades tipicamente requisitadas em sistemas de grande porte. Dentre essas funcionalidades existem várias associadas às questões de segurança, pois essas são vistas como necessidades de primeira ordem nos sistemas supracitados. Nesse sentido, o Grails é uma plataforma que provê vários *plugins* bem reconhecidos na comunidade Java, tais como: SpringSecurity [8], Log4J [9], AuditLogging [10] e JavaMelody [11]. O SpringSecurity provê suporte, por exemplo, para autorização no nível de método e de objetos usando ACLs (*Access Control List*); interfaces básicas para visualizar as configurações de segurança; controle de acesso de múltiplos sistemas independentes (*single sign-on*); autenticação OpenID e LDAP (*Lightweight Directory Access*).

(iii) extensibilidade para viabilizar o acesso dos sistemas via dispositivos móveis, tanto via *browser*, quanto via aplicação nativa. Ambos os casos são suportados pela plataforma Grails. Uma aplicação desenvolvida em Grails, por *default*, pode ser visualizada via *browser* de dispositivos móveis. Testes foram feitos nas arquiteturas móveis Android e FirefoxOS, acessando as aplicações em diferentes *browsers* de internet usando *tablets* e celulares das marcas Samsung® e LG®. O acesso via aplicação nativa é facilitado pelos serviços RESTful [12].

(iv) integração entre sistemas com base no ePING e outros padrões de interoperabilidade do setor público brasileiro. Nesse caso, a ASR especifica o uso de serviços RESTful para integração interna entre sistemas do Órgão, e *WebServices* [12] para integração com sistemas terceiros, de outros Órgãos públicos. Serviço RESTful é um estilo arquitetural que faz uso de vários padrões: HTTP, URI, JSON, XML e outros. O Grails tem suporte que facilita o uso desse tipo de serviço. A abordagem usada na plataforma sugere a construção de serviços RESTful com base em solicitações HTTP, que são gerenciadas pelas controladoras.

3.3 Visão Lógica

A ASR está estruturada de acordo com o Padrão MVC. A arquitetura base é a da plataforma Grails, adaptada de acordo com as necessidades de desenvolvimento de sistemas de informação governamentais brasileiros, conforme apresentado na Figura 01.

Na camada de visão, a ASR comporta os componentes visuais do sistema de informação, ou seja, os componentes de interação direta com o usuário. Nessa camada, a ASR está apoiada no Bootstrap [13]. Trata-se de uma estrutura de *front-end* intuitiva e emergente voltada para o desenvolvimento web padronizado, combinando elementos para alternar o conteúdo visível.

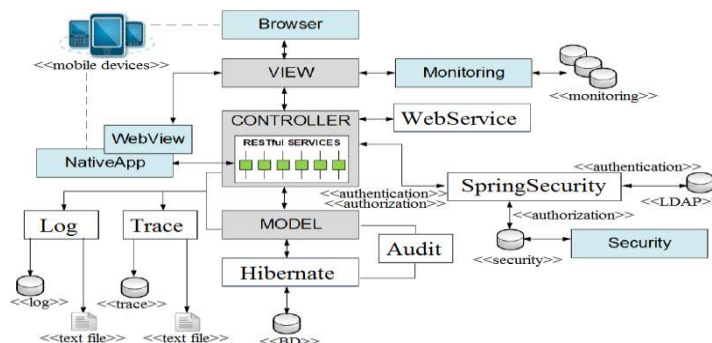


Figura 01: Arquitetura de Software de Referência

Na camada de visão, a ASR comporta os componentes visuais do sistema de informação, ou seja, os componentes de interação direta com o usuário. Nessa camada, a ASR está apoiada no Bootstrap [13]. Trata-se de uma estrutura de *front-end* intuitiva e emergente voltada para o desenvolvimento web padronizado, combinando elementos para alternar o conteúdo visível.

Na camada de controle, os componentes são capazes de receber requisições vindas da visão. Ao enviar uma requisição, a camada de controle trata essa solicitação conforme necessário. Essa camada contém ainda os serviços RESTful. Nesses últimos, as aplicações nativas de dispositivos móveis podem interagir com a aplicação desenvolvida de acordo com as especificações da ASR. Por fim, deve-se considerar que a camada de controle é a responsável por colaborar com a camada de modelo.

Na camada de modelo, têm-se as classes de domínio do sistema de informação, as quais possuem os dados que deverão ser persistidos no banco de dados. Essa camada também é chamada de camada lógica, contendo as regras de negócio da aplicação.

Considerando módulos mais específicos da ASR, destacam-se os módulos responsáveis por: *logging*, *tracing*, monitoramento, auditabilidade, segurança e dispositivos móveis. No Módulo Log, têm-se as funcionalidades para controle do uso de um dado sistema de informação, permitindo o armazenamento de mensagens sobre as requisições que um dado usuário realizou. As mensagens podem ser armazenadas em arquivos de textos, ou mesmo em um banco de dados separado do banco do sistema de informação. Para padronização desse módulo, o mesmo faz uso de um *plugin* específico, chamado Log4J, reconhecido na comunidade de desenvolvedores Java e Grails.

No módulo Trace, esse também é responsável por controlar o uso do sistema de informação, entretanto, focando no controle de erros, falhas e/ou exceções que ocorrerão ao longo da execução desse sistema. Os erros, falhas e exceções podem ser armazenados em arquivos de texto ou outro banco de dados em separado. Esses eventos podem ser lidos normalmente em um arquivo de *logging*, sendo utilizados tipicamente para realização de *debugs*. A própria integração com Log4J já estabelece níveis de *tracing*.

O módulo Monitoring representa um componente separado dos componentes do sistema de informação. Sendo assim, o monitoramento é realizado em outro sistema. Nesse caso, basicamente, o sistema de monitoramento envia uma requisição HTTP ao sistema de informação, verificando o código de retorno dessa requisição. A partir desse código é possível saber se o sistema de informação está funcionando corretamente ou não. Caso não tenha um código de retorno, um erro é apresentado ao usuário. O sistema de monitoramento apresenta os códigos de

retorno em cores, permitindo ao usuário visualizar facilmente o status do sistema de informação monitorado. Os resultados quanto ao monitoramento podem ser salvos em bancos de dados - mais de um. Um *plugin* específico, chamado *JavaMelody*, permite monitorar os sistemas de informação através de gráficos.

No módulo *Audit*, inserções, alterações e remoções são gravadas (i.e. quem fez e quando foi realizado o evento). Tal prática torna o sistema de informação mais seguro e robusto, viabilizando identificar se alguém manipulou o sistema de forma indevida. O *plugin* do Audit Logging, incorporado à arquitetura, oferece suporte completo para auditoria de sistemas.

Existe um componente específico para tratar de segurança. Nesse componente, o *plugin* do Spring Security foi usado para fazer a autenticação e a autorização dos usuários nos sistemas. Trata-se de um *plugin* de código aberto, consolidado na comunidade Java. O Spring Security permite acessar uma base de dados com LDAP, um conjunto de protocolos multiplataforma desenhados para acessar informações distribuídas em uma rede. Adicionalmente, permite realizar a autenticação dos usuários no sistema. Nesse acesso, os dados e papéis de cada usuário são salvos em um banco de dados específico, dedicado às questões de segurança. Têm-se ainda os cadastros das URLs dos sistemas associados, possibilitando definir papéis de acesso para cada URL cadastrada.

A ASR provê também a extensibilidade dos sistemas de informação com dispositivos móveis. Essa extensibilidade pode ser de duas formas: via o próprio *browser* do dispositivo, acessando a URL padrão do sistema; ou por meio de aplicações nativas, instaladas nos próprios dispositivos. Nesse último caso, usam-se serviços RESTful, presentes na *controller*. Ressalta-se que, adicionalmente ao estudo acordado nesse artigo, foram investigadas várias arquiteturas de referência (ex. [14] [15] [16]). Entretanto, tal detalhamento não cabe no presente artigo.

4. PRIMEIRAS IMPRESSÕES

Até o momento, além de várias provas de conceito, conduzidas usando pesquisa-ação e experimentações, a ASR foi testada no desenvolvimento: (i) de um sistema de informação de Gestão, para controle e gestão de projetos por parte da TI do Órgão investigado; (ii) de um sistema de informação para Ouvidoria desse Órgão. Esse encontra-se na primeira *release*; e (iii) de um dos maiores sistemas de informação associados ao Órgão. Esse encontra-se na primeira *release*. Para que o desenvolvimento fosse orientado pela ASR, as terceirizadas de desenvolvimento e de infraestrutura foram capacitadas. Ambientes de desenvolvimento, homologação, teste e produção foram configurados, em parceria com os membros de infraestrutura do Órgão. A ASR demonstrou robustez nos testes de carga e adequação nos testes de vulnerabilidade realizados nos ambientes de homologação do Órgão. Percebeu-se maior produtividade da equipe; satisfação dos clientes, e qualidade no código e nos entregáveis providos com base nas especificações arquiteturais.

5. CONSIDERAÇÕES FINAIS

Foram apresentados os esforços na especificação de uma ASR para novas contratações de desenvolvimento de sistemas de informação governamentais, especialmente desenhado para um Órgão Público Brasileiro. Em resumo, a arquitetura apoia-se: (i) no Paradigma Orientado à Convenção, usando a plataforma Grails, e aumentando a produtividade bem como a flexibilidade e a extensibilidade dos sistemas desenvolvidos; (ii) em serviços

RESTful, visando facilitar a integração com sistemas terceiros, respeitando o ePING; (iii) em *plugins* específicos e reutilizáveis para definição de perfis, *logging* e outras necessidades tipicamente encontradas em sistemas de informação governamentais; (iv) em aplicações específicas para lidar com segurança e monitoramento, o que confere aos sistemas a possibilidade de serem auditados e controlados pelos interessados; (v) em padrões estabelecidos na Engenharia de Software como boas práticas de *design*, implementação, teste e manutenção, no caso: práticas ágeis; código anotado; testes unitários/funcionais/integração, cobertura e outros. Como trabalho em andamento salientam-se esforços na aplicação da ASR em outros sistemas de interesse do Ministério; refinamentos na especificação arquitetural, os quais são pontuais, mas colaboram no ajuste fino de alguns aspectos; e desenvolvimento de *plugins* específicos, visando melhorar ainda mais a reutilização de boas práticas. Como trabalho futuro, pretende-se refinar os módulos da ASR que lidam diretamente com questões de *Web Design*. Nesse sentido, estudos sobre usabilidade, tecnologias e tópicos associados corresponderão aos próximos esforços do grupo.

6. AGRADECIMENTOS

Especiais agradecimentos ao Órgão Público, terceirizadas associadas e demais frentes de pesquisa do projeto, os quais colaboraram na condução da pesquisa apresentada nesse artigo.

7. REFERENCES

- [1] ePING. Acesso: <http://governoeletronico.gov.br/acoes-e-projetos/e-ping-padroes-de-interoperabilidade> (Abril 2015)
- [2] Struts. Acesso: <https://struts.apache.org/> (Abril 2015)
- [3] Miller, J. (2009). *Design For Convention Over Configuration*. Acesso: <http://msdn.microsoft.com/en-us/magazine/dd419655.aspx> (Abril 2015)
- [4] Grails. Acesso: <https://grails.org/> (Abril 2015)
- [5] RubyOnRails. Acesso: <http://rubyonrails.org/> (Abril 2015)
- [6] Play. Acesso: <https://www.playframework.com/> (Abril 2015)
- [7] RebelLabs. (2013) Curious Coder's Java Web Frameworks.
- [8] SpringSecurity, Grails Plugin. Acesso: <http://grails.org/plugin/spring-security-core> (Abril 2015)
- [9] Log4J. Grails Plugin. Acesso: <http://logging.apache.org/log4j/1.2/index.html> (Abril 2015)
- [10] AuditLogging, Grails Plugin. Acesso: <http://grails.org/plugin/audit-logging> (Abril 2015)
- [11] JavaMelody, Grails Plugin. Acesso: <http://grails.org/plugin/grails-melody> (Abril 2015)
- [12] Pautasso et al. (2008). Restful Web Services vs. Big'Web Services. 17th Int. Conf. on World Wide Web (pp. 805-814).
- [13] Bootstrap. Acesso: <http://getbootstrap.com/> (Abril 2015)
- [14] German Federal Government for IT. (2003) Standards and Architectures for e-Government Applications.
- [15] Al-Khanjari et al. (2014) Developing a Service Oriented E-Government Architecture. Journal of Software Engineering and Its Applications. Vol.8(5), pp. 29-42.
- [16] Nakagawa, E. Y.; Oquendo, F.; Becker, M. (2012) RAModel: A Reference Model for Reference Architectures. 6th European Conference on Software Architecture.

Ecosistemas de Startup de Software: Resultados Iniciais no Âmbito do Estado do Pará

Alternative Title: Software Startup Ecosystems: Initial Results in the state of Pará

Nágila Natália de Jesus Torres
Universidade Federal do Pará
Rua Augusto Corrêa, 01 – Guamá.
CEP 66075-110
nagila.ufpa@gmail.com

Cleidson R. B. de Souza
Universidade Federal do Pará &
Instituto Tecnológico Vale
CEP 66055-080
cleidson.desouza@acm.org

RESUMO

O trabalho descrito neste artigo apresenta os resultados iniciais de uma pesquisa mais ampla que visa produzir uma caracterização do ecossistema de *startups* de software no Pará. Esta pesquisa descreverá os diversos atores envolvidos neste ecossistema, bem como as relações entre eles, e através da comparação do ecossistema do Pará com o de outros lugares do Brasil e do mundo, uma análise crítica será feita. A partir disto, poderão ser sugeridas diretrizes e recomendações para o desenvolvimento do ecossistemas de *startups* de software de Belém. Neste trabalho, os resultados iniciais são descritos baseados em um conjunto de entrevistas realizadas com empreendedores localizados em Belém, Pará.

Palavras-chave

Ecossistema, startups, empreendedorismo e inovação.

ABSTRACT

The overall goal of our research is to produce a characterization of the software startup ecosystem in the state of Pará, Brazil. This characterization will describe the various actors involved in this ecosystem, as well as the relationships among them. This paper presents our initial results based on a set of interviews conducted with entrepreneurs located in the city of Belém, Pará. We report different aspects faced by those involved in startup ecosystems, which can be used to suggest actions to strengthen the ecosystem. We plan to collect additional data in other cities of the state as well as to compare our results with other software startup ecosystems.

Categories and Subject Descriptors

D.2.9 [Management]: Programming Teams; K.6.1 [Project and People Management]: Staffing; K.7.2 [Organizations]

General Terms

Management, Design, Economics, Human Factors.

Keywords

Ecosystem, startups, software startups, entrepreneurship, and innovation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.
Copyright SBC 2015.

1. INTRODUÇÃO

Os incentivos e investimentos ao empreendedorismo, assim como a criação de *startups* no Brasil e no mundo, estão sendo cada vez mais alavancados por instituições de fomento. Países como EUA, Israel, Índia e Brasil têm investido em programas dessa natureza com a convicção de que o empreendedorismo é capaz de trazer benefícios como crescimento econômico, geração de renda, atração de fundos de investimentos em tecnologia e inovação e até mesmo a diminuição de desigualdades sociais [1].

No Brasil, exemplos incluem os programas *Startup Brasil* e o *BrasilMaisTI*. No âmbito paraense a Fundação Amazônia Paraense de Amparo a pesquisa (FAPESPA) e o Parque de Ciência e Tecnologia Guamá (PCT Guamá) são instituições que fomentam o empreendedorismo no estado.

As políticas públicas no Brasil têm tido um crescimento exponencial rumo ao desenvolvimento do empreendedorismo voltado para serviços inovadores, que coloca-o como país de destaque no mundo. Entretanto, a atividade empreendedora acontece mais no sentido da necessidade e não dá oportunidade. Além disso, o investimento em Pesquisa, Desenvolvimento e Inovação (PD&I) no Brasil é de 0,55% do PIB, contra 1,87% nos EUA e 2,45% na Coreia do Sul [2]. Complementado este quadro, observa-se que apesar da literatura sobre o empreendedorismo ser abundante [3], existe uma substancial escassez de pesquisas que exploram ecossistemas de startups de software [15]. Dessa forma, surge então a questão que norteia este trabalho: como incentivar, de maneira efetiva, os empreendedores no estado do Pará? Quais políticas públicas devem ser adotadas? Quais são os atores que precisam ser estimulados e com quais estímulos? Quais dificuldades estes empreendedores enfrentam? Em resumo, o que precisa ser feito para que o estado se torne um pólo de empreendedorismo? Os autores deste artigo estão trabalhando em uma pesquisa mais ampla que visa responder a estas perguntas através da identificação do ecossistema de *startups* de software do Pará, bem como das suas relações e de suas limitações. As limitações serão descritas a partir de estudos comparativos com outros ecossistemas. Este trabalho apresenta os resultados iniciais desta pesquisa ao relatar aspectos identificados através de um conjunto de entrevistas com empreendedores de software localizados em Belém, Pará.

2. FUNDAMENTAÇÃO TEÓRICA

Startup é uma empresa dos mais diversos ramos, que surge espontaneamente na condição de extrema incerteza e que tem em sua essência a inovação para criar produtos e serviços [4]. Uma

outra definição de *startups* afirma que as mesmas são organizações temporárias em busca de um modelo de negócio escalável, lucrativo e que possa ser repetido [5]. Em resumo, as distintas definições sobre *startups* concordam que as mesmas são organizações de atividades inovadoras que estão sob os riscos de incertezas a respeito do negócio. Embasada nessa mesma perspectiva Aulet e Murray [6] afirmam que as *startups* pertencem a um tipo de empreendedorismo orientado para a inovação que buscam oportunidades globais e, que esse tipo de empreendedorismo possui uma clara vantagem competitiva e de alto potencial de crescimento.

Nesse contexto, o ecossistema de *startups* é visto como o ambiente mais amplo que afeta o empreendedorismo, principalmente as *startups*. Aleisa [7] define o ecossistema como uma sociedade de empreendedores, ideias, habilidades, startups, incubadoras, capital e os meios de comunicação. Um ecossistema de *startup*, assim como um ecossistema biológico prevê que seus seres vivos convivam mutuamente em uma espécie de dependência natural, assim este também pressupõe que seus elementos interajam de forma harmônica, buscando sempre a adaptação quando algum elemento deixa de existir dentro do ecossistema. A inexistência de um elemento causa um desequilíbrio e quase sempre afeta outro (s) elemento (s). A exemplo disso, Crowne [14] destaca que muitas *startups* de software não alcançam um retorno financeiro sobre os investimentos, e que isso pode ser causado em falhas de venda e execução do produto, desenvolvedores inexperientes, ausência de planejamento estratégico, dentre outros elementos que afetam o ecossistema de *startups* de software.

3. MÉTODO DE PESQUISA

Esta pesquisa será implementada por meio de um estudo qualitativo exploratório, pois esse tipo de abordagem permite estudar os fenômenos que envolvem os seres humanos e suas relações sociais, estabelecida em diversos ambientes de modo a construir hipóteses [9].

Strauss e Corbin [10] classificam três estruturas no estudo qualitativo, são elas: dados, procedimentos e relatórios. Os dados são provenientes de entrevistas, questionários, vídeos e documentos; no procedimento é realizada a interpretação dos dados seguindo de atribuições de conceitos e categorias e, por fim tem-se relatórios, onde são informados os resultados de pesquisa e discussão destes.

Como método de coleta dos dados optou-se pela revisão bibliográfica sobre ecossistemas de *startups* nos mais diversos lugares seguido da condução de entrevistas semiestruturadas [11] com membros do ecossistema paraense.

Para análise dos dados optou-se pela teoria fundamentada em dados [10] que tem por intuito descobrir e organizar conceitos e relações nos dados brutos em um esquema explanatório, isto é, agrupar dados sistematicamente de forma que no final da análise uma teoria baseada nos dados coletados é obtida, pois inicialmente não se sabe qual a teoria, apenas o contexto em que a pesquisa está inclusa.

A coleta e análise de dados deverão ser realizadas de maneira intercalada até a sua saturação teórica [10], ou seja, quando não surgirem mais dados relevantes ou quando estes começarem a se repetir. O que define a próxima etapa de coleta de dados é a amostragem teórica [10], fruto das análises dos dados, em que serão selecionados alguns pontos principais ou pontos que não estão totalmente esclarecidos, de forma a expandir os dados.

4. RESULTADOS INICIAIS

4.1 Coleta de Dados

O início da revisão bibliográfica já foi realizado. Inicialmente, foi feita uma varredura inicial a nível nacional e internacional de materiais que exploram o ecossistema de *startups* dos mais diversos lugares. Os resultados desta pesquisa são brevemente mencionados na seção 5 que compara os ecossistemas identificados com os dados coletados nesta pesquisa.

A primeira coleta de dados foi realizada na cidade de Belém em um evento chamado *Startup Weekend* Belém que ocorreu em Abril de 2014. O evento contou com a presença de cerca de 110 empreendedores. Foram realizadas 18 entrevistas com duração total de 4h e 18 min aproximadamente de áudio, sendo que cada entrevista durou em média 13 minutos. Quando transcritas, as entrevistas resultaram em cerca de 50 páginas de texto. O guia de entrevistas contava com 22 questões abertas.

4.2 Análise dos Dados

Para apoiar a análise qualitativa dos dados utilizou-se a ferramenta MaxQDA versão 2. A Figura 1 ilustra as categorias identificadas nas quais os trechos das entrevistas estão associados.

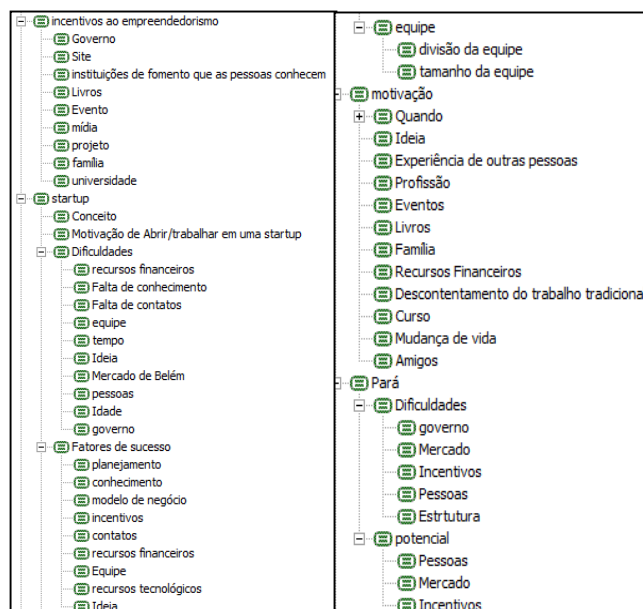


Figura 1. Categorias identificadas

Foi realizada uma análise linha a linha para que as categorias fossem identificadas. As categorias auxiliam o pesquisador a organizar, separar, unir, classificar e validar as respostas encontradas pelos instrumentos de coleta de dados [12].

Após a identificação das categorias, subcategorias foram criadas, a fim de detalhar ainda mais os dados. O objetivo é permitir identificação de correlações com outros dados e a construção de hipóteses que serão confirmadas ou refutadas nas próximas coletas de dados [10].

4.2.1 Categorias Identificadas

Foram identificadas várias categorias no decorrer da análise, contudo apenas 4 classificadas como principais serão apresentadas: motivação ao empreendedorismo, incentivos ao empreendedorismo, dificuldades enfrentadas pelas *startups* e fatores que as *startups* consideram importantes para que as

mesmas tenham sucesso. Abaixo de cada descrição estarão dispostos alguns trechos das entrevistas para confirmar a análise dos dados. Em alguns pontos serão encontradas perguntas de pesquisa (sublinhadas), as quais são hipóteses que irão permitir o surgimento de uma discussão dos resultados.

4.2.1.1 Motivação ao Empreendedorismo

Segundo as entrevistas, um fator motivador é a experiência empreendedora vivenciada por outras pessoas, como por exemplo outros membros da família:

“Familiar sim, sempre foi, da família a gente traz também de educação né? então eu sempre fui muito independente, [...]”

Outro aspecto é que a profissão/emprego motiva as pessoas ao empreendedorismo, principalmente quando o trabalho requer um maior aprendizado sobre gerenciamento, administração, gestão de pessoas e motivação de equipes.

Em contrapartida, há aqueles que se motivaram porque não se adaptam com o modelo de trabalho tradicional, pois este raramente tem flexibilidade de horários e qualidade de vida.

4.2.1.2 Incentivos ao empreendedorismo

A maioria das pessoas conhece pelo menos uma instituição que fomenta o empreendedorismo, mas, segundo os entrevistados, são poucas estas instituições em comparação ao número de brasileiros que querem empreender, além das mesmas serem burocráticas.

“[...] de fomento, tem, sabe ... tem alguns, mas são ... é existe a Universidade Federal do Pará através do centro tecnológico de inovação, eles investem em algumas coisas, existe alguns programas governamentais como o BNDES... Mas burocráticos, todos extremamente burocráticos [...]”

Outra questão é o conhecimento sobre os incentivos ao empreendedorismo destas instituições: a maioria sabe que existe, mas não tem informações sobre como os mesmos funcionam. Hipótese: Será que as pessoas não procuram estas informações, ou estas informações não estão acessíveis?

As universidades possuem um papel importante quanto os incentivos ao empreendedorismo, oferecendo disciplinas de empreendedorismo, administração e o famoso “se vira”, o qual geralmente desafia os alunos a encontrar algum tipo de problema para tentar resolver. Entretanto, segundo eles, atualmente as universidades preparam os alunos para o mercado tradicional e para os concursos públicos e o modelo de ensino atualmente não contempla uma educação voltado ao empreendedorismo.

4.2.1.3 Dificuldades Enfrentadas pelas Startups

A falta de informações sobre incentivos fiscais, burocracia, funcionamento de um negócio, recursos financeiros foi um fator de dificuldades relatados pelos entrevistados.

“Informações que a ideia a gente já tinha é muito boa, mas as informações quanto aos incentivos fiscais, essa coisa de burocracia e tal ninguém tinha [...]”

A equipe da startup é também um fator de dificuldade porque cada sócio pensa de um jeito diferente dos outros sócios e, além disso, outro fator de dificuldade foi que a equipe ao longo do caminho perde motivação, fica desfocada e com outras prioridades. Além disso, o fator tempo também foi levantado porque é difícil conciliar o trabalho na startup, com trabalhos de faculdades, TCC, e outras atividades. Hipótese: Se as pessoas tivessem incentivos financeiros para trabalhar na startup, elas teriam mais tempo em consequência de não ter que trabalhar em outro lugar?

“Ah, acho que te exige muito tempo, pra mim é super difícil conciliar por exemplo, conciliar um evento desse que eu tenho que tá aqui, então não tem como conciliar tudo, mesmo porque eu tô fazendo o meu tcc [...]”

Uma outra dificuldade relatada pelos informantes consiste em “vender” a ideia para as outras pessoas, isto é, é difícil as pessoas acreditarem nas ideias, ou seja, é complicado conseguir a confiança delas de que o produto é aquilo que está anunciando. Hipótese: Essa ideia não é útil para as pessoas ou as ideias estão mal formuladas?

A falta de incentivos do governo foi um fator de dificuldades, pois o governo deveria incentivar a abertura de startups, isenção de impostos e patrocinar eventos de empreendedorismo. Hipótese: a startup foi atrás de incentivos do governo e não encontrou ou ela não procurou?

4.2.1.4 Fatores de Sucesso

Entre os fatores de sucesso podem-se incluir o comprometimento, determinação e persistência da equipe, ou seja, uma equipe disposta é essencial para o sucesso, segundo as entrevistas. Não importa se a pessoa não tenha nível superior, o importante é a capacidade de execução dela. Entretanto outras pessoas acreditam que a capacitação da pessoa principalmente na área da tecnologia é importante.

“Primeiro, pessoas que vão trabalhar contigo na startup, são pessoas que tem que tá dispostas e acessíveis a qualquer situação [...]”

Uma boa ideia também é essencial para a startup ter sucesso. Informação, planejamento, modelo de negócios, incentivos do governo, universidades, investidores, relacionamento com o cliente e um pouco de conhecimento também são fatores importantes. Hipótese: Qual o critério que define se essa ideia é boa ou não?

5. TRABALHOS CORRELATOS

Aleisa [7] realizou uma revisão bibliográfica da literatura sobre os 20 maiores ecossistemas de todo o mundo. O estudo deu enfoque, principalmente, nos ecossistemas do Vale do Silício, Toronto e Moscou, de forma que identificou as principais características desses três ecossistemas de startups e, por fim realizou uma comparação entre eles.

Por outro lado, Kallberg [3] analisou startups brasileiras de base tecnológica, que realizam a maior parte de seus negócios através da internet. O trabalho explorou por meio de entrevistas e questionários os fatores que inibem ou estimulam a produção e o crescimento de startups no Brasil. De maneira similar, Oliveira e colegas [13] exploraram através de pesquisas quantitativas e qualitativas realizadas em 5 estados brasileiros das regiões Sul e Sudeste, as características, forças e fraquezas do ecossistema empreendedor de startups. A pesquisa realizou também um comparativo com ecossistema de outros países.

O trabalho de Kon, Cukier e colegas [15] através de uma ampla pesquisa qualitativa e quantitativa, buscou caracterizar o ecossistema de startup de software de Israel, o que permitiu a construção de um framework com os elementos-chave do ecossistema e suas relações.

Os artigos anteriores concordam que o sistema tributário brasileiro se apresenta como um fator limitante para o crescimento saudável do ecossistema de startups [3, 7, 13], uma vez que o Brasil apresenta taxas elevadas de juros, morosidade em processos e poucos incentivos de retorno ao mercado de empresas falidas.

Estes autores também relatam que há capital de investimento, porém os empreendedores reclamam de não possuírem acesso ao mesmo, pois os investidores argumentam que falta estrutura legal que os incentive a investir em empreendimentos de maior risco. Assim, estes investidores preferem investir em grandes empresas, onde os riscos são menos elevados. Todos estes fatores foram identificados na presente pesquisa na categoria “dificuldades encontradas pelas startups”.

Apesar do brasileiro apresentar resistência em fracassar e aversão ao risco, um ponto positivo é o otimismo do brasileiro em relação a absorção de novos mercado e tecnologias [3, 13]. Este aspecto é mencionado como um fator de sucesso (determinação e persistência) pelos entrevistados. Segundo Oliveira e Colegas [13] isto é motivado pela necessidade de encontrar uma forma de geração de renda em situações onde não existem outras alternativas. O medo do fracasso parece reforçar o perfil de “empreendedor por necessidade” em vez de “empreendedores por oportunidade” [2]. Isso é corroborado na categoria “motivação ao empreendedorismo”, onde os entrevistados afirmam que empreendem por falta de boas oportunidades de emprego e por falta de qualidade de vida. Percebe-se com isso que falta no Brasil, além de estrutura física, uma cultura empreendedora. Segundo Oliveira e colegas [13], a defasagem na educação em relação ao tema empreendedorismo reforça essa característica cultural, pois as instituições de ensino não enfatizam este tema.

De acordo com os quatro estudos mencionados, quanto maior a qualificação dos empreendedores, maiores são as chances de sucesso [3, 7, 13, 15]. Por exemplo, foi percebido que um dos maiores trunfos de Moscou, Sydney e Melbourne e Israel é a qualificação de seus empreendedores. Este dado porém contrasta com dados coletados durante a pesquisa realizada, pois a maioria dos entrevistados respondeu que não acredita ser necessária uma alta qualificação do empreendedor, apenas a sua capacidade de execução. Assim, este resultado sugere que os empreendedores de Belém precisam ser informados sobre a necessidade de qualificação profissional. Além disso, um possível mito sobre os empreendedores de sucesso parece ter sido identificado. Seria importante verificar se este mito sobre a qualificação dos empreendedores ocorre em outras cidades do Brasil.

Em resumo, cidades sedes de ecossistemas de *startups* favoráveis possuem características em comum, tais como: educação voltada para o empreendedorismo, o que justifica na maioria das vezes a presença de uma cultura empreendedora e empreendedores qualificados; incentivos governamentais e qualidade de vida como estímulo para empreender nas cidades.

6. CONCLUSÃO

Este trabalho apresentou os resultados iniciais de uma pesquisa que visa caracterizar o ecossistema paraense de *startups* de software. Estes resultados são baseados em uma análise de um conjunto de entrevistas com empreendedores que relatam pontos positivos e negativos de empreenderem no estado do Pará.

Os próximos passos para o avanço desta pesquisa incluem a realização de etapas adicionais de coleta de dados através de visitas a *startups* de software, incubadoras, universidades, entre outros elementos do ecossistema para a identificação de novas categorias ou até mesmo para a saturação das existentes. Isto será feito em paralelo com uma revisão mais detalhada da literatura. Outro passo importante a ser realizado é a comparação entre o ecossistema paraense com o ecossistema de outras cidades do Brasil e diferentes países.

7. AGRADECIMENTOS

Os autores gostariam de agradecer aos empreendedores pelo tempo dedicado durante as entrevistas, à FAPESPA pela bolsa de mestrado da primeira autora, e finalmente, ao CNPq pelo apoio financeiro através dos processos 485070/2013-8, 440880/2013-0 e 310468/2014-0.

8. REFERÊNCIAS

- [1] Grin, E. J. et al. Desenvolvimento de Políticas Públicas de Fomento ao Empreendedorismo em Estados e Municípios. São Paulo: Programa de Gestão Pública e Cidadania, FGV-EAESP, 2012. 52p.
- [2] Cassapo, F. Inovação no Brasil X Inovação no Mundo. Revista PEGN. 2013. Acessado em Janeiro de 2014 <<http://revistapegn.globo.com/Revista/Common/0,EMI24006618478,00INOVACAO+NO+BRASIL+X+INOVACAO+NO+MUNDO.html>>
- [3] Kallberg, P. An Analysis of Brazil's Startup Ecosystem as a Nurturing Force for Internet-Based Startups: Friend or Foe? 2013, 85f. Trabalho de Conclusão de Curso (Bachelor of Science degree) - Faculty of The School of Business, University of Warwick, 2014.
- [4] Ries, E. A Startup Enxuta. São Paulo: Leya, 2012. 275p.
- [5] Blank, S; Dorf, B. The Startup Owner's Manual: The Step-by-Step Guide for Building a Great Company. Califórnia: K&S Ranch Press, 2012. 557p.
- [6] Aulet, B; Murray, F: A Tale of Two Entrepreneurs: Understanding Differences in the Types of Entrepreneurship in the Economy. Martin Trust Center for MIT Entrepreneurship. 2013
- [7] Aleisa, E. Startup Ecosystems: Study os the ecosystems around the world; focusing on Silicon Valley, Toronto and Moscow. 2013
- [8] Motayama, Y.; Walkins, K. Examining the Connections within the Startup Ecosystem: A CASE Study of St. Louis. Kauffman Foundation Research Series on City, Metro, and Regional Entrepreneurship. 2014
- [9] Godoy, A. S. Introdução à pesquisa qualitativa e suas possibilidades, SP, v. 35, n. 2, p.57-63, Marc/Abr. 1995
- [10] Strauss, A.; Corbin, J. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. 2008.
- [11] Manzini, E. J. A entrevista na pesquisa social. São Paulo: Didática, 1991. 158p.
- [12] Bartelmebs, R. C. S/A. Metodologias de Estudos e Pesquisas em Educação III. Analisando os dados na pesquisa qualitativa. Recuperado em 24 de Julho de 2014. <http://repositorio.sead.furg.br:8080/jspui/bitstream/1/1453/1/Texto_analise.pdf>
- [13] Oliveira, A. A. et al. O Ecossistema Empreendedor Brasileiro de Startups: Uma análise dos determinantes do empreendedorismo no Brasil a partir dos pilares da OCDE. Núcleo de Inovação e Empreendedorismo, 2013.
- [14] Crowne, M. Why software product startups fail and what to do about it., 2000.
- [15] Kon, F.; Cukier, D.; et al. A Panorama of the Israeli Software Startup Ecosystem. 2014.

Uma Arquitetura de Referência para Plataforma de Crowdsensing em Smart Cities

Alternative Title: A Reference Architecture for a Crowdsensing Platform in Smart Cities

Herbertt B. M. Diniz
Centro de Informática, UFPE
Jornalista Aníbal Fernandes
Recife, Brazil 50.740-560
hbmd@cin.ufpe.br

Emanoel C. G. F. Silva
Centro de Informática, UFPE
Jornalista Aníbal Fernandes
Recife, Brazil 50.740-560
ecgfs@cin.ufpe.br

Kiev Santos da Gama
Centro de Informática, UFPE
Jornalista Aníbal Fernandes
Recife, Brazil 50.740-560
kiev@cin.ufpe.br

RESUMO

Em decorrência dos problemas gerados pelo crescimento populacional nas grandes cidades, surge a necessidade de soluções que apontem para a iniciativa de “Cidades Inteligentes” (*Smart Cities*), ou seja, que utilizam a tecnologia para oferecer recursos, que podem ajudar a solucionar ou minimizar os problemas urbanos. Essas soluções buscam a integração de diversas fontes de Tecnologias da Informação (TICs), no entanto, essas fontes de Tecnologias da Informação formam estruturas complexas e geram um grande volume de dados, que apresentam enormes desafios e oportunidades, dificultando a possibilidade de se disponibilizar programas que integrem informações de sensores e que captem o espaço físico, tirando amostras do que está acontecendo em tempo real. Com intuito de fornecer uma arquitetura de referência, a fim de confrontar questionamentos acerca desses desafios e oportunidades, neste artigo é apresentada uma abordagem, utilizando componentes *off-the-shelves* (componentes de prateleira), uma plataforma de *Crowdsensing* para solução de *smart cities*, também foi realizado um experimento para determinar a resistência e a estabilidade do sistema. Assim, esta proposta abre caminho para que se amplie a integração de dados de fontes dos mais variados tipos de sensores e dispositivos.

Palavras-Chave

Cidades Inteligentes, CEP, Arquitetura, Sensores Humanos

ABSTRACT

Due to the problems caused by population growth in large cities, there is a need for solutions that point to the initiative of Smart Cities, that is, using the technology to offer resources that can help solve or minimize urban problems. This solutions seeks the integration of several sources of In-

formation Technology (ICTs), however, those source of Information Technology form complex structures and generate a large volume of data, that represents big challenges and opportunities, impeding the possibility of to make available, systems that integrate informations of sensors and capture data from the physical space, getting samples of what is going on in the city in real time. With the intention of offer a reference architecture, to compare issues related with those challenges and opportunities, this paper presents an approach that employs components off-the-shelves for the construction of a crowdsensing platform for solution in Smart Cities. We performed an experiment to determine the performance and stability of the system. Thus, this proposal opens the way for to broaden the integration of data sources of various types of sensors and devices.

Categorias e Descritores do Assunto

D.2.11 [Software Engineering]: Software Architectures;
D.2.12 [Software Engineering]: Interoperability

Termos Gerais

Performance Standardization

Keywords

Smart Cities, CEP, Architecture, Crowdsensing

1. INTRODUÇÃO

Em consequência da evolução tecnológica, vários dispositivos eletrônicos, informatizados, estão se integrando através de redes de internet, se incorporando a todos os espaços e ambientes das cidades. Isso permite a utilização de tecnologias para monitoramento urbano[20], influenciando a nossa maneira de interagir com os espaços urbanos os quais estão cada vez mais caóticos, em decorrência do inchaço populacional.

Dados estatísticos demonstram que há um enorme crescimento populacional nas grandes cidades, que tende a aumentar de forma alarmante nos próximos anos[6], acarretando assim, enormes desafios na solução de problemas legados a vários domínios específicos, principalmente, no que diz respeito à mobilidade urbana, segurança, educação, utilização eficiente dos recursos naturais e descarte de resíduos, sem que haja prejuízos para o meio ambiente.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

Em decorrência desses fatos, surge a necessidade de soluções que convirjam para iniciativa de “Cidades Inteligentes” (*smart cities*), ou seja, aquelas que utilizam a tecnologia para oferecer recursos que podem ajudar a solucionar ou minimizar os problemas urbanos em seus domínios específicos. Cidades inteligentes[19] é um termo que ganhou força na academia, empresas e governo para descrever cidades que, por um lado, são cada vez mais compostas e monitoradas por computação pervasiva e ubíqua e, por outro, cuja economia e governança está sendo impulsionada pela inovação, criatividade e empreendedorismo, interpretados por especialistas.

Segundo Naphade *et al.*[19], várias iniciativas já estão em curso nesse sentido, a exemplo do modelo de sustentabilidade aplicado em Dubuque no estado americano Iowa, em Seul na Coreia do Sul e do Centro de Operações da prefeitura do Rio de Janeiro. Elas buscam através da integração de diversas fontes de Tecnologias da Informação (TICs), tais como Internet das Coisas (IoT), Cloud Computing, Sistemas de Informações Geográficas (SIGs) e Dispositivos Móveis, além da utilização dos próprios cidadãos, levando em consideração uma visão holística da cidade e centralizando o monitoramento e controle para auxiliar no apoio à tomada de decisão [15].

No entanto, essas fontes de TICs formam estruturas complexas[18], e geram um grande volume de dados, que apresentam enormes desafios e oportunidades, dificultando a possibilidade de se disponibilizar programas que integrem informações de sensores e que captem o espaço físico, tirando amostras do que está acontecendo em tempo real. A maioria das soluções existentes nesse sentido são proprietárias e possuem alto custo, a despeito disso o uso combinado de FLOSS (*Freely Licensed Open Source Software*)[22] e COTS (*Commercial Off-The-Shelf*)[9] possibilita a criação de soluções de baixo custo e com pouco esforço de implementação, segundo Gasperoni[16] o uso de FLOSS e COTS oferece inúmeras vantagens, a exemplo da qualidade de software e rapidez na evolução.

Com intuito de fornecer uma arquitetura de referência, a fim de confrontar questionamentos acerca desses desafios e oportunidades, neste artigo é apresentada uma abordagem, utilizando componentes *off-the-shelves* (componentes de prateleira), que atende e implementa de forma eficaz e simplificada uma plataforma de *middleware* para solução de *smart cities*, configurável para vários domínios, utilizando Processamento de Eventos Complexos, para análise de streams de dados em tempo real e integrando sensores através de serviços web e fila de mensagem assíncrona.

Foram conduzidos alguns experimentos para provar a viabilidade da abordagem proposta. Assim, foram realizados testes para avaliar a capacidade de resposta às requisições, eficiência do tempo de resposta e o comportamento do serviço assíncrono de fila de mensagem. O objetivo dos testes foi determinar a resistência e a estabilidade do sistema.

O presente trabalho está organizado da seguinte forma: na Seção 2, são apresentados trabalhos relacionados; na Seção 3, é apresentada a visão geral e descrição do sistema; na Seção 4, são apresentados as ferramentas e os métodos dos testes realizados na Seção 5. Na Seção 6, são apresentados resultados e discussões e, por fim, na Seção 7, são apresentadas conclusões e trabalhos futuros.

2. TRABALHOS RELACIONADOS

Muitas pesquisas relacionadas ao gerenciamento e integração de sensores heterogêneos em espaços públicos, estão baseadas em *mobile crowd sensing* e EDA (*Event Driven Architecture*). Segundo Etzion e Niblett[12], EDA é um estilo arquitetural em que um ou mais componentes de software respondem ao receber um ou mais eventos de notificação. *Mobile Crowdsensing*[15], por sua vez, refere-se a sensoramento comunitário (*i.e.*, através de pessoas) para coleta de dados, no qual as pessoas funcionam como sensores (*crowd sensors*). Abaixo é apresentada uma lista não exaustiva de plataformas construídas utilizando as abordagens citadas.

SOFIA[14] é uma plataforma de interoperabilidade semântica entre um conjunto selecionado de aplicações e um ecossistema entre objetos interativos, como sensores, atuadores e outros sistemas embarcados. Seu objetivo é oferecer recursos para a construção de ambientes inteligentes. Alguns requisitos destes ambientes inteligentes são o armazenamento de informação e procedimentos de busca a sistemas embarcados, implementadas em diferentes tecnologias. O projeto aborda três áreas de aplicação que representam diferentes tipos de espaço em termos de escala, aplicações e serviços: espaços pessoais inteligentes (por exemplo, automóveis), espaços interiores inteligentes (por exemplo, casa e escritório), e espaços da cidade inteligentes (por exemplo, infra-estrutura alargada e facilidades como uma estação de metrô, shopping e assim por diante).

Borja e Gama[8] propõem uma plataforma baseada em barramento de serviços com o objetivo de integrar diferentes fontes de dados em diferentes sistemas e de dispositivos. Por exemplo prático, a plataforma é aplicada ao domínio de monitoramento do sistema de transporte público da cidade do Recife. Como resultado, foi demonstrado que aplicações de terceiros podem ser construídas a partir dos dados coletados destas fontes de dados heterogêneas.

TrafficInfo[13] é um aplicativo baseado em *Crowd sensors*, que visualiza em tempo real as informações de transporte público da cidade pelo Google Maps. Ele é construído utilizando modelo *publish/subscribe* para comunicação dos passageiros que se inscrevem no TrafficInfo, de acordo com seu interesse, para os canais de informação de trânsito dedicadas a diferentes linhas de transportes públicos ou pontos de paradas. Assim, eles são informados sobre a situação do transporte público ao vivo, como por exemplo, as posições reais dos veículos. TrafficInfo basicamente visualiza a circulação de veículos de transporte público em um mapa. Quando os dados em tempo real são recolhidos pelos passageiros, a informação da posição adquirida é atualizada refletindo a situação do tráfego. Assim, os passageiros podem ter uma ideia se o veículo selecionado está na hora ou atrasado, se eles têm que se apressarem ou procurarem algum transporte alternativo.

CrowdOut[7] é também um serviço de *crowd sensing* móvel para cidades inteligentes, que permite aos usuários relatar problemas de segurança rodoviária que eles experimentam em seu ambiente urbano (excesso de velocidade, carros estacionados ilegalmente, sinais e símbolos, qualidade de estradas e níveis de tráfego, etc.) e partilhar esta informação com a sua comunidade. Basicamente, um usuário pode especificar o tipo de crime, tirar uma foto com seu smartphone e adicioná-lo ao relatório para apresentá-la; ele também pode adicionar um comentário curto (140 caracteres, como no Twitter) para descrever a infração de forma mais precisa. O serviço também registra as coordenadas

GPS dos usuários, a fim de indicar com precisão o local do delito ocorrido na estrada. Em seguida, os relatórios são exibidos num mapa em tempo real.

Choi & Kang[10] propõem uma plataforma em HTML5, com serviço baseado em localização, fazendo uso do GPS, a fim de achar pontos de interesse nas proximidades. Como resultado, apresentam um case de uma ferramenta de dispositivo móvel, para ajudar estrangeiros e coreanos num complicado processo de registro de aquisição e alienação de imóveis na Coreia do Sul. Neste trabalho foi desenvolvido um aplicativo móvel utilizando o framework de aplicação Sencha Touch[5] utilizando serviços RESTful através do framework Restlet[4].

Ao analisar estas soluções, percebe-se que a solução proposta neste trabalho se diferencia da plataforma SOFIA por possuir um escopo mais fechado, objetivando oferecer serviços de *Crowdsensing*. Em contrapartida, a apresentada pela SOFIA aborda várias áreas de aplicações. Apesar disso, nossa proposta permite a configuração de domínio, contrastando nesse aspecto com as plataformas TrackInfo e a proposta apresentada por Choi & Kang. A plataforma CrowdOut, contrasta com a nossa por não possuir recurso de informação de alertas disparados através de processamento de eventos complexos. Já Borja e Gama propõem uma solução utilizando arquitetura de barramento que, pelo fato de oferecer recursos além do necessário para fornecer os serviços propostos, torna a solução mais complexa e difícil de administrar.

3. VISÃO GERAL E DESCRIÇÃO DO SISTEMA

3.1 Visão Geral

Diante das soluções de *middleware* para *smart cities* propostas mais comumente atualmente, algumas características foram levadas em consideração na construção de nossa arquitetura, por serem tratadas de forma comum nas mais diferentes abordagens. A exemplo disso, pode-se citar o uso de conexão de sensores usando serviços de fila de mensagem para a integração de fontes de dados de maneira assíncrona e síncrona utilizando serviço web restful, processamento de eventos complexos para análise de *streams* de dados em tempo real e uso de sensores humanos através de sistemas de *Crowdsensing*. Para tal, utilizamos alguns componentes de prateleira, com intuito de facilitar e simplificar a implementação. O papel desempenhado por cada componente será apresentado mais adiante na descrição da arquitetura da plataforma, a qual foi construída baseada nessas características.

O projeto arquitetural da plataforma é descrito em 2 módulos de serviços: *Front-End Module* e *Back-End Module*, conforme mostra a Figura 1. No Módulo *Front-End* (Figura 1 (1)), encontra-se o sistema de *Crowdsensing*, no módulo de *Back-End* (Figura 1 (2)), a API de serviços restful (Figura 1 (2.a)), que coordena o acesso aos Serviços Síncronos de *Request/Response* e ao serviço de Processamento Eventos (Figura 1 (2.c)). Este, por sua vez, acessa os serviços do módulo serviço de mensagens assíncronas no modelo *Publish/Subscribe* (Figura 1 (2.d)), disponibilizando assim a transmissão de dados em tempo real e a sinalização de dispositivos, que possibilita estabelecer e manter conexões via Socket e Restful, permitindo qualquer dispositivo enviar e

receber dados para serem processados e redistribuídos. Estes dados são controlados pelo Gerenciador de Serviços da aplicação, (Figura 1 (2.b)), que acessa através do serviço de persistência, os dados persistidos na base de Dados (Figura 1 (2.e)).

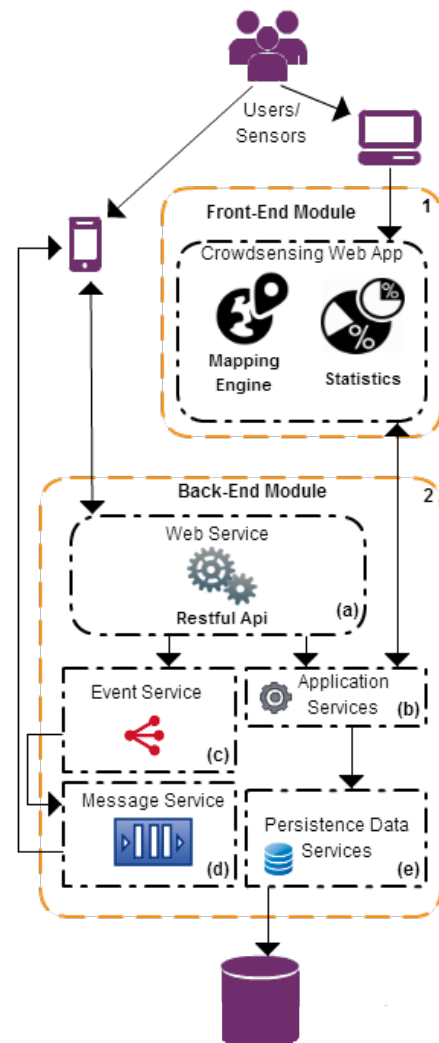


Figura 1: Arquitetura da plataforma

3.2 Descrição do Sistema

A Plataforma possui dois serviços de entrada de dados, que podem ser providos através de Sistema de *Crowdsensing* no módulo *Front-End* ou através da API Restful do *Web Service* no módulo *Back-End*. A seguir descrevemos respectivamente as duas abordagens.

3.2.1 Sistema de Crowdsensing

A plataforma proposta possui um recurso que permite a participação colaborativa dos próprios usuários, possibilitando a captação da percepção dos mesmos sobre o ambiente em que vivem, tornando possível a inserção das pessoas como elementos mais participativos no sistema e na cidade, transformando-os em sensores humanos (*Crowd Sensors*). Segundo Erickson[11], “humanos podem contribuir com co-

nhecimento qualitativo mais profundo, analisar dados vagos ou incompletos e agir de uma forma em que sistemas digitais frequentemente não são capazes". Nesse âmbito foi desenvolvido um Sistema de *Crowdsensing*, que possui um serviço do Google Maps, o qual permite aos usuários na forma de sensores humanos visualizar pontos de ocorrências no mapa através de marcadores, os gráficos com estatísticas e a tabela com a lista de detalhada com dados de ocorrências.

Segundo Sherchan *et al.*[21], há duas categorias de *crowdsensing*: *personal* e *community sensing* (sensoriamento pessoal e comunitário), onde o sensoriamento pessoal refere-se ao tratamento de dados de um único indivíduo para beneficiar a este mesmo indivíduo com o resultado do processamento dos dados por ele emitido. Em se tratando de sensoriamento comunitário, [15] existem mais dois tipos de modelos distintos: o sensoriamento oportunista, que requer um envolvimento mínimo do usuário (ex: localização automática por GPS), e o sensoriamento participativo, que requer a participação ativa e direta no fornecimento de dados dos usuários.

Na plataforma proposta, a aplicação de *crowdsensing* é parametrizável para vários domínios oferecendo a oportunidade de criar marcadores customizáveis, a exemplo do mapa descrito na Figura 2, com dados do domínio de mobilidade urbana, descrevendo pontos de acidentes de trânsito ocorridos com bicicletas. Essa ferramenta se enquadra na categoria de *community sensing* e no modelo de sensoriamento participativo, por ser voltada para o controle de fenômenos de grande escala, e de muitos indivíduos participantes e pela participação direta dos usuários.

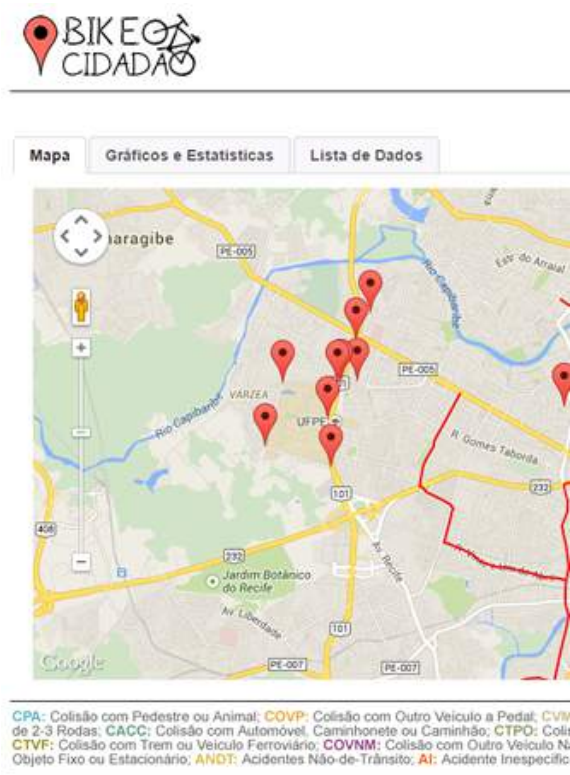


Figura 2: Tela principal do sistema de *crowdSensing*

A ferramenta web ainda possui um painel com gráficos e

estatísticas que apresenta a totalização de ocorrências, um gráfico em formato pizza que mostra a estatística de cada tipo de ocorrência, e a lista de ocorrências em formato tabular.

3.2.2 Services

Segundo Maréchaux[17], aplicações de negócios precisam ser conectadas a fim de criar uma solução integrada. Arquitetura Orientada a Serviços (*Service-oriented Architecture*-SOA) e arquitetura orientada a eventos (*Event Driven Architecture*-EDA) são dois paradigmas diferentes que abordam os desafios da integração complexos. Essas aplicações tem como características principais flexibilidade e responsividade. O desafio de TI tem sido apoiar esta visão de negócios com as arquiteturas e tecnologias apropriadas. O modelo a ser descrito a seguir é uma proposta simplificada baseada na utilização de EDA e SOA através do uso de alguns componentes de prateleira.

O modelo de *Web Services*, apresentado na Figura 3, provê serviços de dados em tempo real através de requisições síncronas e assíncronas. As requisições síncronas são feitas através de web service RESTful utilizando JSON como padrão de transmissão de dados. As requisições via REST são realizadas através do envio de dados do cliente solicitando serviços providos pela camada *Web Service* via POST/GET que, por sua vez, solicita dados da persistência através de funções parametrizáveis. Como exemplo, tem-se o serviço de localização que retorna o tipo de ocorrência e a distância entre um ponto de latitude e longitude enviado pelo cliente e a localização do ponto de ocorrência mais próximo, persistido pelo usuário(sensor) através do serviço de *Crowdsensing*, descrito no Item 3.2.1

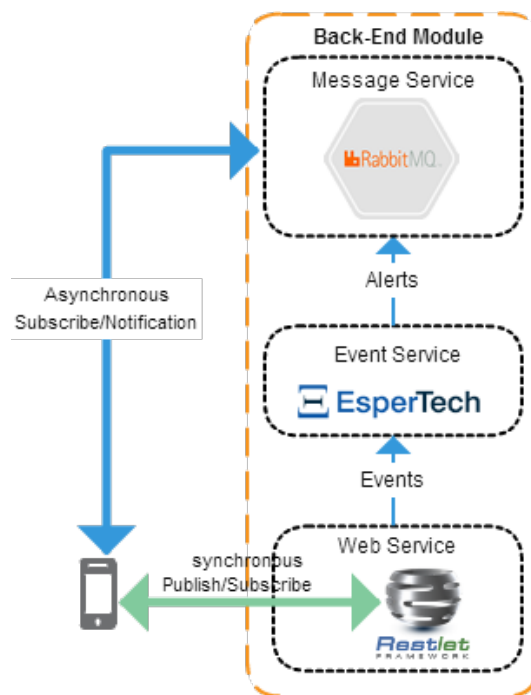


Figura 3: Modelo da camada *Service*

À medida que as requisições chegam, eventos são registrados no sistema CEP Esper[2] e alertas são disparados para

filas no serviço de processamento de eventos do modelo assíncrono para serem consumidos pelos clientes. Desse modo, esse modelo utiliza o serviço de mensagens em fila, para capturar os alertas de CEP disparados de maneira assíncrona, através da captura de eventos e processamento das regras, além de fornecer serviços através do framework Restlet, processando ocorrências armazenadas na camada de dados, enviadas pelos sensores humanos.

O Módulo de *Web Services* provê serviços de localização que podem ser parametrizáveis através das funções e das customizações das regras da ferramenta de CEP, além de possibilitar o usuário a criação de novas regras para serem providas através do serviço de filas do sistema RabbitMQ[3], no modelo assíncrono. Uma das regras desse modelo (ilustrada pelo Listing 1) é de localização parametrizável, que permite disparar alertas com a distância, em tempo real, das últimas duas requisições de localização de pontos de latitude e longitude enviados através do serviço de REST e inseri-las na fila de mensagens para serem consumidas pelos consumidores registrados.

Listing 1: Exemplo de regra

```
SELECT 6378137 * Math.sqrt(
    Math.pow((Math.toRadians(cast(Vehicle.lat,
    double)) - Math.toRadians(cast(
    BikePosition.lat, double))), 2) +
    Math.pow((Math.toRadians(cast(Vehicle.
    lng, double)) - Math.toRadians(cast(
    BikePosition.lng, double))), 2))
AS distance, bikePosition.imei,
    bikePosition.lat, bikePosition.lng,
    Vehicle.imei, Vehicle.lat, Vehicle.lng
FROM BikePosition.win:length(1)
AS bikePosition, Vehicle
ORDER BY distance ASC LIMIT 1
```

Foi desenvolvida uma aplicação Android para o case de acidentes de trânsito ocorridos com bicicletas, com intuito de consumir dados da plataforma e apresentar informações sensíveis ao contexto. O aplicativo possui duas funcionalidades básicas, providas através do consumo de dados do Módulo de *Web Services*. Um desses serviços é o “Alerta de *Crowdsensing*”, onde o aplicativo envia sua posição de GPS atual e recebe a informação da distância entre o ciclista e os pontos de acidentes registrados pelo sistema de *Crowdsensing*, alertando ao ciclista que ele está próximo a um ponto de acidente. Outro serviço é o “Alerta de CEP”, que em tempo real pode identificar a aproximação de algum usuário que esteja usando o aplicativo no modo veículo motorizado, e ser alertado a fim de evitar possíveis colisões.

4. AMBIENTE DE TESTE

Esta seção discute os materiais e métodos utilizados para a prova de conceito da plataforma proposta. Para a execução dos testes foram utilizados dois computadores, cuja configuração está descrita na tabela 1.

Tabela 1: Configuração das máquinas utilizadas

	Processador	Frequência	RAM
M.USER	2410M	2,3GHz	DDR3 8GB
M.SERVER	M380	2,53GHz	DDR3 4GB

O computador chamado M.USER foi utilizado para simular usuários e enviar status e requisições à plataforma. A simulação das requisições foi realizada em M.USER utilizando a ferramenta de testes de performance JMeter¹ versão 2.9. O computador chamado M.SERVER, foi aquele que hospedou a plataforma, recebeu e processou as requisições de M.USER.

Ao todo foram realizados dois testes. O primeiro teste (a partir daqui chamado de T1) simulou apenas requisições onde os usuários não passavam por pontos de acidentes, previamente cadastrados pelo serviço de *Crowdsensing*. No segundo teste (a partir daqui chamado de T2), em todas as requisições os usuários estavam em rota de perigo, segundo o serviço de *Crowdsensing*, e estavam em zona de colisão com outro usuário. A Seção 5 descreverá o resultado de cada teste.

Considerando a natureza em tempo real dos dados trafegados pela plataforma, dada a necessidade de manter atualizada a posição geográfica dos usuários, para a configuração dos usuários virtuais no JMeter foram definidos os seguintes parâmetros: *Threads* = 1000, *Ramp-up* = 2 e *LoopCount* = *forever*. Ou seja, indefinidamente, a cada 2 segundos 1000 usuários virtuais enviam requisições à plataforma. Em cada teste o JMeter aferiu o desempenho da plataforma recebendo requisições simultâneas dos 1000 usuários virtuais durante 15 minutos. A performance foi monitorada pelas seguintes métricas: *Times Over Time*, *Transactions per Second*, *Response Times Percentiles*, *Time Graph*.

As requisições dos usuários virtuais foram do tipo GET e foram configuradas com os seguintes parâmetros: *Connect Timeout* = 10000 (ms), *Response Timeout* = 10000 (ms). O *Path* de conexão para envio das requisições segue a segue o seguinte formato: */project/rest/trace/[id da fila]/[latitude]/[longitude]/*, como por exemplo: */project/rest/trace/54321/-8.05529673777147/-34.951613545417786/*.

Além disso, em M.SERVER foi feita uma configuração para lidar com a grande quantidade de requisições dos usuários virtuais e diminuir o custo para a abertura de conexões. Através da biblioteca C3PO² do hibernate³ foi configurado um *pool* de 1000 conexões.

5. RESULTADOS DOS TESTES

Nesta seção serão consolidados os resultados dos dois testes realizados.

Como apresentado anteriormente, no teste T1 foi simulado um total de 1000 usuários enviando seu status ao servidor e esperando alguma resposta sobre a situação da rota em curso. Neste teste, 100% dos usuários estavam em rota segura, ou seja, não haviam notificações do serviço de *Crowdsensing*. Portanto, nenhum evento foi disparado pelo serviço de CEP e nenhuma mensagem foi adicionada à fila de mensagens do usuário no RabbitMQ. Por outro lado, no teste T2 100% dos usuários estavam em uma rota que tinha alguma notificação pelo serviço de *Crowdsensing*. Dessa forma, a cada requisição, um evento CEP era disparado e uma mensagem era adicionada à fila de mensagens do usuário com a descrição do acidente logo à frente.

Dessa forma, o teste T1 possui uma característica exclu-

¹Disponível em <http://jmeter-plugins.org>

²Disponível em <http://www.mchange.com/projects/c3p0/>

³Disponível em <http://hibernate.org/>

sivamente síncrona. Neste teste através de um *web service RESTful* provido pela plataforma o usuário envia sua coordenada geográfica e apenas espera um documento JSON com a distância e as características do ponto de acidente mais próximo. De fato, nenhum usuário neste teste estava em rota de risco. O Teste T2, por sua vez, possui características síncrona e assíncrona. Além da conexão síncrona do teste T1, o teste T2 calcula a distância do ciclista em relação ao automóvel mais próximo. Em caso de ambos estarem abaixo de uma distância tolerável o serviço de eventos disparava alertas nas filas de mensagens do ciclista e do automóvel. Dessa forma, ambos mantinham uma conexão assíncrona com a plataforma para serem notificados da proximidade de outro usuário.

5.1 Tempo de resposta

O objetivo desta métrica foi verificar o tempo médio que a resposta a uma requisição demora para chegar até o usuário virtual. Ou seja, o tempo que inicia quando a requisição é enviada ao servidor e que termina quando a resposta é completamente recebida pelo cliente. No JMeter é representada pela métrica *Times Over Time* e *Time Graph*.

Como esperado, em geral o tempo de resposta de T2 foi superior ao tempo de T1. Como ilustrado pela Figura 4, o tempo médio de resposta para T1 foi de 3708,58 ms e 4660,02 ms para T2. De fato, tomando 30 amostras aleatórias de tamanho 100, com reposição nos tempos de respostas mensurados, e considerando um nível de 95% de confiança, tem-se um intervalo de tempo médio de resposta de 3604.47ms a 3840.0ms para o teste T1 e de 4509.861ms a 4627.493ms para o teste T2.

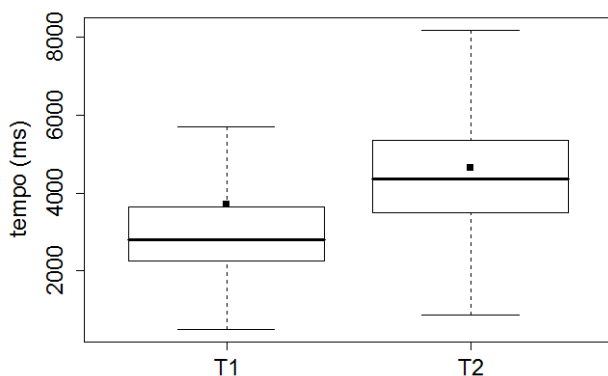


Figura 4: Distribuição dos tempos de resposta. Dada a variação apresentada nas medições de tempo, sua representação ficou melhor detalhada em um gráfico em caixa. A linha preta, como padrão, representa a mediana e o quadrado em preto representa a média.

Ao contrário de T1, em T2 há o tempo dedicado à execução do módulo de processamento de eventos complexos e de notificações nas conexões assíncronas com o RabbitMQ. Através das médias elencadas anteriormente, pode-se calcular que o tempo gasto com essas tarefas é de, em média, 951,44 ms. Ou seja, esse é o tempo médio que a plataforma leva para receber um fluxo de coordenadas geográficas, calcular a distância desta em relação aos pontos de acidentes, verificar se este ponto está em uma zona de risco e, por fim,

disparar notificações nas filas de mensagens do RabbitMQ.

Para entender melhor o tempo de resposta, a Figura 5 mostra o percentil dos tempos de resposta de cada um dos testes. Pode-se observar que o tempo médio de resposta de T1 está no percentil 79,0%. Ou seja, 79,0% das requisições tem um tempo menor ou igual à média. Da mesma forma o tempo médio de resposta de T2 está no percentil 68,8% (ver Figura 5). No JMeter o percentil é dado pela métrica *Response Times Percentiles*.

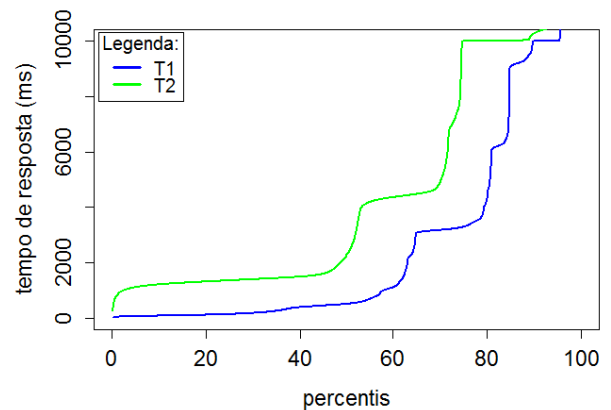


Figura 5: Percentil do tempo de resposta. Os valores de percentil ordenam crescentemente a amostra em 100 partes. Dessa forma, o n -ésimo percentil representa os primeiros $n\%$ dos valores da amostra.

5.2 Taxa de sucesso e falha

O objetivo desta métrica foi evidenciar o comportamento de sucesso e de falha nas requisições nos teste T1 e T2. Como mencionado na Seção 4, as requisições do JMeter foram configuradas com os parâmetros de *Connect Timeout* = 10000 (ms) e *Response Timeout* = 10000 (ms). Ou seja, caso uma conexão demorasse mais que dez segundos para ser aberta ou caso a requisição demorasse mais que dez segundos para ser respondida, esta requisição era considerada falha. No JMeter é representada pela métrica *Transactions per Second*.

Observando a Figura 6, pode-se notar que para o teste T1 houve uma média de 313,31 requisições por segundo processadas com sucesso e 21,02 requisições por segundo falhas. Já para o teste T2 houve uma média de 189,14 requisições por segundo concluídas com sucesso e 32,53 requisições por segundo falhas. De fato, assim como na Seção 5.1, tomando 30 amostras de tamanho 100 dos tempos de respostas mensurados e considerando um nível de confiança de 95%, tem-se um intervalo de médio de 306.78 a 317.32 requisições bem sucedidas por segundo no teste T1 e 186.93 a 191.29 para o teste T2. Já para as falhas, os intervalos de confiança são de 19.77 a 21.34 requisições falhas no teste T1 e 31.22 a 33.28 falhas para o teste T2.

No teste T2, para cada requisição respondida com sucesso, duas mensagens eram enviadas ao RabbitMQ, sendo uma para a fila de mensagens do ciclista e outra à fila de mensagens do automóvel. Acessando no servidor o endereço `localhost:15672/#/`, é possível encontrar o painel de controle do RabbitMQ e acompanhar em tempo real as atividades do serviço de fila de mensagem. Ao todo foram enviadas

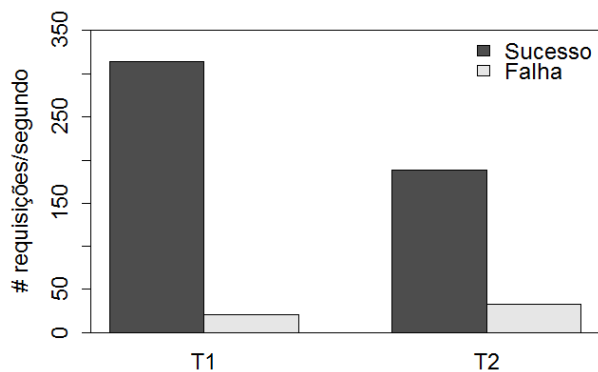


Figura 6: Taxa de sucesso e de falha nas requisições. As barras escuras representam as requisições concluídas com sucesso. As barras claras representam as requisições falhas.

350.307 mensagens ao RabbitMQ, correspondendo a uma média de 389,23 mensagens por segundo (ver Figura 7). De fato, esta taxa é praticamente o dobro da quantidade de requisições concluídas com sucesso em T2.



Figura 7: Taxa mensagens publicadas na fila do RabbitMQ por segundo. O período ilustrado corresponde aos últimos dez minutos de teste.

6. RESULTADOS E DISCUSSÃO

Como ilustrado na Seção 5, considerando um total de 1000 requisições simultâneas a plataforma apresentou um tempo médio de resposta de 3708,58 ms para rota segura (teste T1) e um tempo médio de 4660,02 ms para notificação de proximidade com outros usuários (teste T2). Sendo que, para T1, 79% das requisições tem um tempo menor ou igual à média e para T2, 68,9%. Além disso, de uma média de 334,33 requisições por segundo, para T1, apenas 6,28% delas não foram concluídas com sucesso. E para uma média de 221,67 requisições por segundo em T2, 14,67% foram requisições falhas.

Em um confronto dos resultados obtidos com uma demanda real de requisições, pode-se usar como exemplo o projeto BikePE[1]. Nesta iniciativa, 800 bicicletas estão disponíveis para aluguel em 80 estações espalhadas pela Região Metropolitana de Recife. A ideia é que através de uma opção sustentável se possa melhorar a mobilidade urbana e ao mesmo tempo oferecer mais opções de lazer à população.

Segundo BikePE[1], de segunda a sexta-feira há uma demanda média de 1100 viagens. Ou seja, apesar de diferentes contextos de aplicabilidade, e considerando os dados de desempenho apresentados pela plataforma pode-se perceber que utilizando apenas um computador com configurações

domésticas seria o suficiente para suprir a demanda de um projeto desta dimensão em uma cidade como Recife utilizando a solução proposta neste trabalho.

Além disso, em caso de haver uma maior demanda de requisições, a plataforma pode ser instanciada em diferentes máquinas, oferecendo então a possibilidade de escala horizontal. Ou seja, para atender a uma demanda cada vez maior, não é necessário melhorar o hardware da máquina que hospeda a plataforma. Na verdade, instâncias da plataforma podem ser replicadas em várias máquinas com configurações tradicionais e assim obter uma maior escala a um menor custo.

7. CONCLUSÃO E TRABALHOS FUTUROS

Este artigo apresentou a proposta e a validação de uma plataforma que, ao fazer uso de componentes *off-the-shelves*, atende e implementa de forma eficaz e simplificada uma plataforma de middleware para solução de *Smart Cities*. Dentre suas características, estão: ser configurável para vários domínios, utilizar Processamento de Eventos Complexos para análise de *streams* de dados em tempo real, integrar sensores através de serviços Restful e de fila de mensagem e utilização de sensores humanos através do uso de *crowdsensing*.

A partir da análise dos resultados, pode-se concluir que a plataforma comportou-se de maneira satisfatória nos testes realizados, uma vez que suportou uma carga de requisições semelhantes às demandas reais, considerando que a máquina que hospedou a plataforma tinha a configuração típica de um computador doméstico.

Como ameaças à validação, tem-se o fato de não ter sido utilizada uma rede de dados 3G para a comunicação entre os usuários virtuais e a plataforma. No entanto, como os testes tiveram o objetivo de provar a possibilidade de desenvolvimento da plataforma, não foi necessário fazer testes de latência de rede, por exemplo. Dessa forma, preferiu-se optar pela simulação em uma rede *wireless* local.

De fato, testes de laboratório não reproduzem em sua totalidade o ambiente real de produção. Evidencia-se então a necessidade de realização de um estudo de caso com aplicação a acidentes de trânsito ocorridos com bicicletas. O objetivo é avaliar com maior precisão a eficiência da plataforma, analisando fatores como a latência de rede e pontos de saturação.

Por fim, o presente trabalho prova o conceito de que é possível construir uma plataforma de *middleware* para *smart cities*, baseada em localização, utilizando processamento de eventos complexos, parametrizáveis para vários domínios e capaz de processar regras elaboradas pelos próprios usuários.

Como trabalhos futuros pretende-se ampliar a integração de dados de fontes dos mais variados tipos de sensores e dispositivos, para vários tipos de protocolos de comunicação, além de ampliar a visualização de funcionalidades através do painel de dados estatísticos, apresentando dados em tempo real.

8. AGRADECIMENTOS

Este trabalho foi parcialmente financiado pelo Instituto Nacional de Ciência e Tecnologia para Engenharia de Software (INES⁴), CNPq e FACEPE⁵ sob o processo nº IBPG-0773-1.03/13.

⁴<http://www.ines.org.br>

⁵<http://www.facepe.br>

9. REFERÊNCIAS

- [1] Bikepe. <http://www.bikepe.com/>. Acessado em 1 Feb de 2015.
- [2] Espertech. <http://www.espertech.com>. Acessado em 1 Jan de 2015.
- [3] Rabbitmq, clustering guide. <http://www.rabbitmq.com/clustering.html>. Acessado em 1 Jan de 2015.
- [4] Restlet - restful web framework for java. <http://www.restlet.org/>. Acessado em 1 Jan de 2015.
- [5] Sencha touch - html5 mobile framework. <http://www.sencha.com/products/touch/>. Acessado em 1 Jan de 2015.
- [6] United nations. world urbanization prospects: the 2011 revision. ny. Acessado em 1 Jan de 2015.
- [7] E. Aubry, T. Silverston, A. Lahmadi, and O. Festor. Crowdout: A mobile crowdsourcing service for road safety in digital cities. In *Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2014 *IEEE International Conference on*, pages 86–91. IEEE, 2014.
- [8] R. Borja and K. Gama. Middleware para cidades inteligentes baseado em um barramento de serviços. In *X Simpósio Brasileiro de Sistemas de Informação (SBSI)*, volume 1, pages 584–590, 2014.
- [9] L. Brownsword, T. Oberndorf, and C. A. Sledge. Developing new processes for cots-based systems. *IEEE Software*, 17(4):48–55, 2000.
- [10] M. Choi and B.-T. Kang. Html5 platform independent mobile application using location-based service. *Life Science Journal*, 11(7), 2014.
- [11] T. Erickson. Geocentric crowdsourcing and smarter cities: Enabling urban intelligence in cities and regions. In *1st Ubiquitous Crowdsourcing Workshop at UbiComp*, 2010.
- [12] O. Etzion and P. Niblett. *Event Processing in Action*. Manning Publications Co., Greenwich, CT, USA, 1st edition, 2010.
- [13] K. Farkas, A. Z. Nagy, T. Tomás, and R. Szabo. Participatory sensing based real-time public transport information service. In *Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2014 *IEEE International Conference on*, pages 141–144. IEEE, 2014.
- [14] L. Filippini, A. Vitaletti, G. Landi, V. Memeo, G. Laura, and P. Pucci. Smart city: An event driven architecture for monitoring public spaces with heterogeneous sensors. In *Sensor Technologies and Applications (SENSORCOMM)*, 2010 *Fourth International Conference on*, pages 281–286. IEEE, 2010.
- [15] R. K. Ganti, F. Ye, and H. Lei. Mobile crowdsensing: current state and future challenges. *Communications Magazine, IEEE*, 49(11):32–39, 2011.
- [16] F. Gasperoni. Floss, cots, and safety: A business perspective. *3rd European Congress ERTS – Embedded Real Time Software*, 2006.
- [17] J.-L. Maréchaux. Combining service-oriented architecture and event-driven architecture using an enterprise service bus. *IBM Developer Works*, pages 1269–1275, 2006.
- [18] T. Nagel, M. Maitan, E. Duval, A. V. Moore, J. Klerkx, K. Kloeckl, and C. Ratti. Touching transport-a case study on visualizing metropolitan public transit on interactive tabletops. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, pages 281–288. ACM, 2014.
- [19] M. Naphade, G. Banavar, C. Harrison, J. Paraszczak, and R. Morris. Smarter cities and their innovation challenges. *Computer*, 44(6):32–39, 2011.
- [20] N. Odendaal. Information and communication technology and local governance: understanding the difference between cities in developed and emerging economies. *Computers, Environment and Urban Systems*, 27(6):585–607, 2003.
- [21] W. Sherchan, P. P. Jayaraman, S. Krishnaswamy, A. Zaslavsky, S. Loke, and A. Sinha. Using on-the-move mining for mobile crowdsensing. In *Mobile Data Management (MDM)*, 2012 *IEEE 13th International Conference on*, pages 115–124. IEEE, 2012.
- [22] D. A. Wheeler. Why open source softwarefree software (ossfs, floss, or foss)? look at the numbers! http://www.dwheeler.com/oss_fs_why.html. Acessado em 1 Jan de 2015.

Predição de Rankings de Artistas por Meio de Regressão

Alternative Title: Prediction of Artists' Rankings by Regression

Felipe L. M. Faria
Departamento de Ciência da
Computação
Universidade Federal
de Ouro Preto
Ouro Preto - MG - Brasil
felipelmfaria@hotmail.com

Álvaro R. Pereira Jr.
Departamento de Ciência da
Computação
Universidade Federal
de Ouro Preto
Ouro Preto - MG - Brasil
alvaro@iceb.ufop.br

Luiz H. C. Merschmann
Departamento de Ciência da
Computação
Universidade Federal
de Ouro Preto
Ouro Preto - MG - Brasil
luizhenrique@iceb.ufop.br

RESUMO

A construção de *rankings* consiste na ordenação de resultados recuperados de acordo com um determinado critério. Os *rankings* podem fornecer informações relevantes para analistas de diferentes setores da indústria. Na indústria fonográfica, os *rankings* possibilitam a compreensão de como os estilos musicais e a popularidade dos artistas e suas músicas evoluem com o tempo, permitindo análises de históricos de execuções e de tendências. Devido à importância da construção de *rankings* no escopo musical, técnicas de mineração de dados têm sido utilizadas para prever os *rankings* através de informações contidas em mídias sociais. Este trabalho avalia modelos de regressão para a predição de *rankings* de artistas utilizando-se de dados históricos (*rankings* diários de artistas) extraídos do website Vagalume. Três técnicas de regressão (k-Nearest Neighbors - k-NN, Regressão Linear Múltipla - RLM e Random Forests - RF) foram avaliadas neste trabalho considerando-se diversos cenários. Os resultados obtidos por meio de experimentos mostraram que predições com baixas taxas de erros podem ser obtidas, indicando que técnicas de mineração de dados podem ser utilizadas para obtenção de informações que auxiliem a indústria fonográfica na tomada de decisões.

Palavras-Chave

Mineração de Dados, Regressão, Mídia Social

ABSTRACT

The construction of rankings consists of ordering retrieved results according to certain criteria. Rankings can provide relevant information to analysts from different sectors of industry. For the music industry, rankings enable understanding how musical genres and popularity of artists and their songs evolve over time, allowing analyses of history data and trends. Due to the importance of building rankings in the musical scope, data mining techniques have been used

to predict rankings by using information from social media. This work evaluates regression models for prediction of artists' rankings using historical data (daily rankings of artists) extracted from website Vagalume. Three regression techniques (k-Nearest Neighbors - k-NN, Multiple Linear Regression - MLR and Random Forests - RF) were evaluated in this study considering different scenarios. Results obtained from experiments showed that predictions with low error rates can be obtained, indicating that data mining techniques can be used to obtain information to assist the music industry in decision making.

Categories and Subject Descriptors

H.2.8 [Database Applications]: Data mining

General Terms

Experimentation

Keywords

Data Mining, Regression, Social Media

1. INTRODUÇÃO

Segundo os autores do trabalho [11], a construção de *ranking* consiste na ordenação de resultados recuperados de acordo com um determinado critério. No contexto fonográfico, a construção de *rankings* possibilita a compreensão de como os artistas, estilos musicais e a própria música evoluem ao longo do tempo, podendo nos auxiliar no entendimento do processo criativo de fazer música [3]. Além disso, os *rankings* podem auxiliar críticos de música a analisarem tendências a partir do posicionamento dos artistas nos mesmos.

O volume de vendas de músicas está perdendo seu papel como indicador de popularidade de artistas. Embora os números de vendas e execuções em rádios possam ter sido muito utilizados para medir a popularidade de um artista nas décadas de 50 e 60, essas variáveis tornaram-se cada vez mais pobres como indicadores de popularidade em detrimento ao rápido crescimento da divulgação da música em meio digital [9]. No meio digital, as mídias sociais têm sido muito utilizadas para geração de *rankings* de artistas. Um exemplo disso é o *ranking* "Social 50" divulgado na revista Billboard, o qual é construído a partir de dados coletados em mídias sociais.

Com o aumento da quantidade de informação disponível na *web*, a realização de *rankings* tornou-se um problema

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

importante a ser estudado [11]. Desse modo, no contexto musical, diversas técnicas de mineração de dados têm sido utilizadas para predição de *rankings* de artistas. Essas técnicas usam dados históricos para gerar modelos que posteriormente são utilizados para inferir novos *rankings*.

O Vagalume¹, um *website* colaborativo brasileiro de música, é uma mídia social bastante popular. De acordo com o *ranking* divulgado no *website* Alexa², que leva em consideração o número de acessos a um *website*, o Vagalume tem se mantido ao longo do tempo bem classificado quando comparado com outras mídias brasileiras, como o Letras³, o Kboing⁴ e a versão brasileira do LastFM⁵.

O Vagalume nos permite ouvir músicas, visualizar notícias do meio musical, criar *playlists* (listas de músicas), enviar letras, responder enquetes, comentar sobre os artistas, seguir o que outro usuário está fazendo nessa mídia social, entre outras atividades. Além disso, ele disponibiliza um *ranking* diário de artistas, o qual é construído a partir da popularidade do artista no *website*.

Assim, este trabalho utiliza os dados históricos (*rankings* diários de artistas) fornecidos pelo *website* Vagalume para avaliar a capacidade de modelos de regressão na predição de *rankings* de artistas. Três técnicas (*k-Nearest Neighbors* – *k-NN*, Regressão Linear Múltipla – RLM e Random Forests – RF) foram utilizadas para a construção dos modelos de regressão avaliados. A geração de modelos precisos é importante para auxiliar a indústria fonográfica brasileira nas análises de tendências, por exemplo, indicando quais artistas estarão no topo do *ranking* numa data futura.

As demais seções deste trabalho estão organizadas como descrito a seguir. A Seção 2 contém um breve resumo de alguns trabalhos relacionados. Em seguida, a metodologia utilizada neste trabalho é apresentada na Seção 3. A descrição dos experimentos e os resultados obtidos são mostrados na Seção 4. Por fim, a Seção 5 apresenta as conclusões e sugestão de trabalho futuro.

2. TRABALHOS RELACIONADOS

Os autores do trabalho [9] ressaltam a importância da utilização das mídias sociais para a classificação de artistas por meio da construção de *rankings*. Assim, devido à importância desse assunto, vários autores têm realizado esforços na geração de *rankings* confiáveis e precisos para atender diversas necessidades de usuários do meio fonográfico.

No trabalho [5], os autores propuseram uma abordagem, que utiliza técnicas de mineração de dados em texto, para medir a popularidade e construir *rankings* de artistas a partir da análise dos comentários dos ouvintes na rede social MySpace⁶, uma mídia musical popular. Eles demonstraram ter obtido resultados mais próximos àqueles que seriam gerados manualmente por usuários (estudantes universitários), do que os que são gerados a partir dos métodos utilizados pela revista Billboard.

Autores do trabalho [3] construíram um *ranking* de artistas com base no *website* WhoSampled⁷, um *website* de in-

formações sobre o meio musical. Para isso, utilizaram várias características de artistas e gêneros musicais e adotaram a métrica PageRank para permitir a interpretação e descrição de padrões de influência musical, tendências e características de músicas.

Indo além de métodos que utilizam as mídias digitais para a construção de *rankings*, outros trabalhos já foram realizados para auxiliar o meio fonográfico em suas diversas necessidades. Nesses trabalhos, a utilização de técnicas de regressão no auxílio à construção de *rankings* relacionados ao meio fonográfico tem gerado resultados satisfatórios. Alguns desses trabalhos são apresentados a seguir.

No trabalho [14], os autores construíram um sistema de reconhecimento de emoções que utiliza valores de valência (polaridade das emoções) e de excitação (intensidade das emoções) para recomendar músicas para usuários. De acordo com a emoção escolhida por um usuário, o sistema retorna músicas caracterizadas com valores de valência e excitação mais próximos àqueles associados à emoção escolhida pelo usuário. As emoções são tratadas como variáveis contínuas cujos valores são computados por meio de funções que combinam valores de valência e de excitação. Técnicas de regressão foram utilizadas para predizer os valores de valência e de excitação das músicas a partir de características (timbre do cantor, intensidade dos agudos, sonoridade) dos segmentos das mesmas. Dentre os resultados obtidos, foi constatado que realizar a predição utilizando técnicas de regressão é melhor do que realizá-la a partir de técnicas de classificação, onde haveria a dificuldade para se distinguir entre as emoções “feliz” e “satisfeito”. Além disso, os autores afirmaram que os resultados de acurácia preditiva obtidos superaram aqueles apresentados na literatura.

Os autores do trabalho [4] propuseram a recomendação de artistas a partir das características musicais dos artistas e das preferências dos usuários. Nesse trabalho, um artista foi caracterizado por parâmetros do seu modelo acústico correspondente. A partir disso, esses modelos acústicos juntamente com os *rankings* de preferência dos usuários foram utilizados como vetores de entrada para gerar um modelo de regressão que posteriormente foi usado para recomendar outros artistas para os usuários. Os resultados experimentais desse trabalho demonstraram que essa abordagem superou outras em termos da qualidade da recomendação.

No trabalho [13], os autores propuseram classificar músicas por meio de características denominadas de conceito, que podem ser gêneros, instrumentação, características vocais e outros. Para isso, treinaram um modelo de conceitos através de uma técnica de regressão ordinal, que foi utilizada visando minimizar o problema de desbalanceamento de classes da base de dados utilizada nos experimentos. Com essa abordagem, em uma avaliação que foi realizada sobre um conjunto de dados, os autores mostraram que foram capazes de melhorar a precisão da detecção de conceitos em 10,89% em média em relação a outros trabalhos apresentados na literatura.

Diferentemente do foco dos trabalhos descritos anteriormente, neste trabalho utilizam-se técnicas de regressão para predizer as posições dos artistas no *ranking* de uma determinada data a partir de suas posições em datas anteriores por meio de dados obtidos em uma mídia social. Espera-se, que assim como nos trabalhos citados anteriormente, a utilização da tarefa de regressão gere resultados satisfatórios.

¹<http://www.vagalume.com.br>

²<http://www.alexa.com>

³<http://www.letras.mus.br>

⁴<http://www.kboing.com.br>

⁵<http://www.lastfm.com.br>

⁶<http://www.myspace.com>

⁷<http://www.whosampled.com>

3. METODOLOGIA

Nesta seção, é apresentada a maneira como as bases de dados foram construídas, na Seção 3.1. Em seguida, na Seção 3.2, é realizado um estudo dessas bases de dados.

3.1 Bases de Dados

O *website* Vagalume foi a mídia social escolhida para a coleta dos dados utilizados na construção das bases de dados adotadas neste trabalho. Esse *website* disponibiliza *rankings* de artistas que são construídos de acordo com a popularidade deles nessa mídia levando-se em consideração a quantidade de acessos realizados por usuários. Para cada artista, a rede social disponibiliza um histórico contendo sua posição diária no *ranking* ao longo do tempo. Esse histórico da posição dos artistas no *ranking* foi coletado utilizando-se uma API (*Application Programming Interface*) disponibilizada pelo Vagalume. A título de ilustração desse histórico, a Figura 1 apresenta um gráfico disponibilizado pelo *website*. Nesse gráfico, o eixo *x* representa a data e o eixo *y* a posição do artista no *ranking*.



Fonte: Adaptado do *website* Vagalume

Figura 1: Exemplo do *Ranking* diário de um artista no *website* Vagalume.

As bases de dados utilizadas neste estudo foram construídas escolhendo-se 478 artistas de 63 gêneros diferentes de música. Como há mais de 478 artistas no *ranking* gerado pelo Vagalume, para garantir que os artistas selecionados fossem classificados entre as posições 1° e 478°, uma reclassificação foi realizada respeitando-se a ordem da posição dos artistas no *ranking* original do Vagalume. O período considerado para coleta de dados foi de 13 de janeiro a 03 de abril de 2014.

A seguir, a Figura 2 apresenta a estrutura das bases de dados construídas. Cada instância das bases representa um artista que é caracterizado pelas suas posições nos *rankings* gerados em diferentes dias. Com o objetivo de avaliar a predição em diferentes datas, os dias 21 (classe 21), 41 (classe 41) e 81 (classe 81) foram escolhidos como atributo classe para as diferentes bases de dados geradas. Além disso, para cada atributo classe, quatro bases de dados foram construídas variando-se a quantidade de dias, os atributos preditores. Foram construídas bases com a quantidade de 10, 20, 40 e 80 atributos preditores.

Bases com as características mencionadas anteriormente foram geradas para diferentes quantidades de artistas de acordo com o gênero musical dos mesmos, a saber: Sertanejo, Rock, Pop, MPB e todos os gêneros juntos. Sendo assim, foram construídas bases contendo somente artistas do gênero Sertanejo (41 instâncias), somente artistas do gênero Rock (134 instâncias), somente artistas do gênero Pop (109

Artistas/Dias	Dia 1	Dia 2	Dia 3	...	Dia n
Artista A	82	82	96	...	...
Artista B	83	87	81	...	...
Artista C	84	84	86	...	...
...	...	...	...	...	...
Artista Z	86	90	92	...	...

Figura 2: Exemplo da base de dados construída do *Ranking* diário de artistas.

instâncias), somente artistas do gênero MPB (63 instâncias) e o total de artistas de todos os gêneros (478 instâncias).

A construção dessas bases considerando-se diferentes gêneros foi realizada com intuito de se avaliar diferentes cenários. Os gêneros Pop e Sertanejo representam os gêneros mais acessados na *web*. Em 2013, por exemplo, segundo a lista das expressões mais acessadas do Google⁸, o gênero Pop foi um dos mais buscados, e no Youtube Brasil⁹, dentre os vídeos mais acessados, estão os gêneros Pop e Sertanejo. Pesquisa realizada pela empresa Crowley¹⁰, especializada em monitoração eletrônica de *broadcast* de áudio no Brasil desde 1997, aponta que nos últimos 14 anos os gêneros Sertanejo e Pop seguem em alta, sendo o primeiro e o segundo mais executados ao longo dos anos nas rádios, respectivamente. Por outro lado, o gênero MPB e Rock vêm decaindo em termos do número de execuções em rádios.

3.2 Análise de Dados

Um estudo nas bases construídas foi realizado com o intuito de se observar como a posição dos artistas no *ranking* varia ao longo do tempo. Desse modo, o gráfico da Figura 3 apresenta o cálculo da média da variação de posições para os artistas classificados em faixas do *ranking* para uma das bases (base com 10 atributos preditores e a classe 81). As faixas são: 1° ao 10°, 11° ao 20°, 21° ao 30° e assim por diante até o intervalo 471° ao 478°. O gráfico em questão mostra essa variação média de posições para cada um dos intervalos mencionados anteriormente considerando os artistas de todos os gêneros musicais. Vale ressaltar que para todas as bases construídas, o comportamento dos gráficos foi semelhante. No gráfico da Figura 3, o eixo *x* apresenta os intervalos de posições de 10 em 10 e o eixo *y* mostra a variação média de posições.

No cenário apresentado no gráfico da Figura 3 pode-se observar que para alguns intervalos as variações de posições em dias subsequentes são menores, indicando que a posição dos artistas no *ranking* é mais estável ao longo do tempo. Por outro lado, intervalos com variações maiores indicam uma maior oscilação de posições ao longo dos dias. Por exemplo, no gráfico da Figura 3 pode-se verificar que a variação média de posições no *ranking* para artistas classificados entre 421° e 430° é igual a 10. Isso significa que, entre dias subsequentes, a posição de um artista classificado nesse intervalo varia em média 10 posições. De modo geral, essa análise mostrou que as variações menores estão associadas aos intervalos das posições iniciais (entre 1° e 50°) e finais do *ranking* (entre 431° e 478°), enquanto as maiores variações ocorrem nos intervalos que representam as posições intermediárias (entre

⁸<http://www.google.com/trends/topcharts#geo=BR&date=2013>

⁹<http://www.youtube.com/playlist?list=PLoeZWzNXmy8FAz3OUQln9CTWKq5lCs5i>

¹⁰<http://www.crowley.com.br>



Figura 3: Variação média de posições considerando-se 10 atributos preditores e a classe 81.

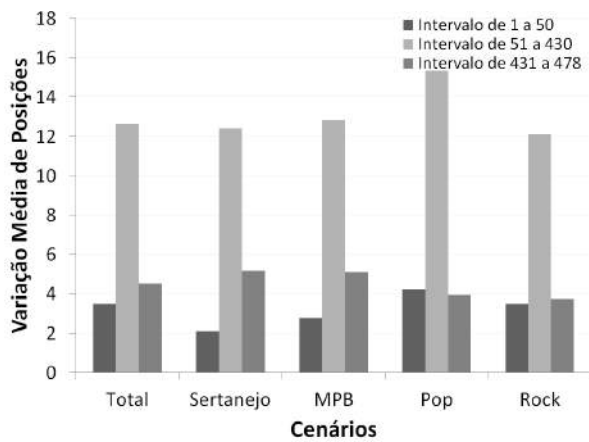


Figura 4: Variação média de posições considerando-se 10 atributos preditores e a classe 81 para os artistas de todos os gêneros e os artistas dos gêneros MPB, Pop, Rock e Sertanejo separadamente.

51° e 430°) do ranking.

O padrão de variação de posições observado para os artistas presentes separadamente nos gêneros MPB, Pop, Rock e Sertanejo é semelhante àquele apresentado no gráfico da Figura 3 (para os artistas de todos os gêneros musicais). Isso pode ser observado no gráfico da Figura 4, que apresenta para todos os cenários estudados a variação média de posições dos artistas no ranking ao longo do tempo para três intervalos de posições: intervalo das posições iniciais (entre 1° e 50°), intervalo das posições intermediárias (entre 51° e 430°) e intervalo das posições finais do ranking (entre 431° e 478°). Nesse gráfico, o eixo x apresenta os diferentes cenários e o eixo y mostra a variação média de posições.

Essa análise será útil para ajudar a explicar os resultados obtidos a partir dos modelos de regressão, os quais serão apresentados na Seção 4.2.

4. EXPERIMENTOS COMPUTACIONAIS

Nesta seção, será apresentada a configuração experimental, mostrando as técnicas e as métricas utilizadas, na Se-

ção 4.1. Em seguida, na Seção 4.2 serão apresentados os resultados experimentais obtidos no estudo realizado neste trabalho.

4.1 Configuração Experimental

Para os estudos conduzidos neste trabalho foram escolhidas três técnicas de regressão bastante difundidas na literatura: k -NN (k -Nearest Neighbors) [1], RF (Random Forests) [2] e RLM (Regressão Linear Múltipla) [7]. Os experimentos foram realizados utilizando-se as implementações contidas no *software* R [10] (para as técnicas k -NN e RF) e na ferramenta Weka [6] (para a RLM).

Para a técnica k -NN, avaliou-se modelos de regressão com os seguintes valores do parâmetro k : 1, 3, 5, 7 e 10. Para a técnica RF dois parâmetros foram ajustados: para o n tree, que define o número de árvores da floresta, foram avaliados os valores 500 e 1000 (recomendado em [2]); para o m try, que define o número de atributos que serão considerados na construção de cada árvore, foram avaliados os seguintes valores: o número de atributos preditores divididos por 3, 1/6 da quantidade de atributos preditores e 2/3 da quantidade de atributos preditores (sugeridos em [2]).

Para avaliar a precisão preditiva dos modelos de regressão utilizou-se a técnica de validação cruzada denominada *Leave-One-Out* [8]. Nessa técnica, considerando-se uma base de dados históricos com n instâncias, n modelos de regressão são construídos a partir de bases de treinamento e testados a partir de bases de teste. Para isso, a base de dados históricos é subdividida em treinamento e teste n vezes. Em cada subdivisão, das n instâncias que compõem a base de dados históricos, $n - 1$ são utilizadas na formação da base de dados de treinamento e a instância restante compõe a base de dados de teste. Vale ressaltar que cada uma dessas n instâncias da base de dados históricos é utilizada uma vez para compor a base de teste. Para medir a acurácia preditiva dos modelos utilizou-se a métrica Desvio Médio Absoluto (DMA), que computa a diferença entre os valores reais e os valores preditos. A Equação 1 apresenta o cálculo do DMA.

$$DMA = \frac{\sum_{t=1}^n |e_t|}{n}, \quad (1)$$

onde e_t é a diferença entre os valores real e predito e n é o

número de instâncias da base.

Além do DMA, outra métrica de avaliação foi utilizada: o Coeficiente de Determinação (R^2) [12]. O R^2 representa o poder explicativo do modelo de regressão, que tem por objetivo avaliar o quanto os atributos preditores conseguem explicar o atributo classe. Quanto mais próximo o R^2 for de 100% melhor é o modelo, enquanto que resultados próximos a 30% são considerados fracos.

4.2 Resultados Experimentais

Apesar de diversos resultados terem sido gerados a partir de diferentes valores de parâmetros das técnicas de regressão avaliadas (ver Seção 4.1), apenas os melhores resultados serão apresentados a seguir. Esses resultados foram gerados com os seguintes valores de parâmetros: k igual a 3 (para a técnica k -NN), $ntree$ igual a 1000 e $mtry$ igual a 2/3 da quantidade de atributos preditores (para a técnica RF).

As Tabelas 1, 2, 3, 4 e 5 apresentam os resultados de DMA e R^2 para os diferentes cenários avaliados. Nessas tabelas, a primeira coluna refere-se à quantidade de atributos preditores (10, 20, 40 e 80 atributos); a segunda coluna contém as classes utilizadas (classe 21, classe 41 e classe 81) e as três colunas seguintes contêm os resultados (DMA e R^2) das três técnicas avaliadas (k -NN, RF e RLM). Os melhores e piores valores de DMA estão em negrito. Vale observar que para determinadas combinações de “Quant. Atributos Preditores” e “Atributo Classe” os resultados estão marcados com o símbolo –, indicando que para aquele atributo classe não existia na base de dados históricos aquela quantidade de atributos preditores para geração do modelo de regressão.

A Tabela 1 apresenta os resultados de DMA para a base que contém artistas de todos os gêneros (478 instâncias). Dentre os diversos cenários avaliados (variando-se a quantidade de atributos preditores e a classe a ser predita), obteve-se melhor resultado para a classe 81, em que a aplicação da técnica RLM na base com 10 atributos preditores resultou num DMA de 7,5. O pior resultado foi obtido para a classe 41, em que a técnica k -NN para a base com 40 atributos preditores apresentou um DMA de 15,41. Comparando as três técnicas, os melhores resultados de DMA sempre foram obtidos pela técnica RLM e os piores pelo k -NN, permanecendo o RF com desempenho intermediário entre a RLM e o k -NN.

Tabela 1: DMA considerando todos os artistas

Quant. Atributos Preditores	Atributo Classe	k -NN (DMA/ R^2)	RF (DMA/ R^2)	RLM (DMA/ R^2)
10	classe 21	14,18/0,970	12,25/0,977	11,11/0,979
	classe 41	13,69/0,978	12,55/0,981	11,18/0,984
	classe 81	10,21/0,986	8,96/0,989	7,50/0,991
20	classe 21	14,49/0,968	12,26/0,977	11,50/0,978
	classe 41	14,69/0,974	12,59/0,981	10,61/0,984
	classe 81	11,39/0,982	8,89/0,990	7,61/0,991
40	classe 21	–	–	–
	classe 41	15,41/0,967	12,51/0,992	11,61/0,980
	classe 81	12,36/0,981	9,00/0,989	8,05/0,991
80	classe 21	–	–	–
	classe 41	–	–	–
	classe 81	13,56/0,975	9,08/0,989	8,78/0,989

A Tabela 2 apresenta os resultados de DMA para os artistas do gênero Sertanejo. Novamente o melhor resultado

(10,84 de valor de DMA) foi obtido para a classe 81 utilizando-se a técnica RLM na base com 10 atributos preditores. O pior resultado (85,96 de DMA) foi obtido para a classe 41 com a técnica RLM na base com 40 atributos preditores. De modo semelhante à análise anterior, na maioria dos casos, a técnica RLM alcançou os melhores resultados e o k -NN os piores, ficando o RF com os resultados intermediários.

Tabela 2: DMA considerando o gênero Sertanejo

Quant. Atributos Preditores	Atributo Classe	k -NN (DMA/ R^2)	RF (DMA/ R^2)	RLM (DMA/ R^2)
10	classe 21	19,79/0,957	18,79/0,965	15,70/0,973
	classe 41	16,78/0,972	15,76/0,977	11,02/0,988
	classe 81	15,02/0,975	13,10/0,982	10,84/0,985
20	classe 21	19,54/0,966	19,95/0,966	22,81/0,949
	classe 41	16,85/0,970	15,98/0,974	14,40/0,989
	classe 81	13,87/0,980	12,13/0,981	13,48/0,978
40	classe 21	–	–	–
	classe 41	16,60/0,970	16,31/0,972	85,96/0,266
	classe 81	14,15/0,979	14,33/0,973	47,67/0,798
80	classe 21	–	–	–
	classe 41	–	–	–
	classe 81	16,19/0,964	17,78/0,961	12,61/0,977

A Tabela 3 apresenta os resultados de DMA para os artistas do gênero Pop. O melhor resultado (9,19 de DMA) foi obtido pela técnica RLM para a classe 21 utilizando-se a base com 10 atributos preditores. O pior resultado (DMA de 25,89) foi encontrado para a classe 41 com a técnica RLM para a base que contém 40 atributos preditores. Para os artistas do gênero Pop, na maioria das vezes, a RLM supera as duas outras técnicas para as bases contendo 10 e 20 atributos preditores. Já para as bases com 40 e 80 atributos preditores, a técnica RF supera as demais.

Tabela 3: DMA considerando o gênero Pop

Quant. Atributos Preditores	Atributo Classe	k -NN (DMA/ R^2)	RF (DMA/ R^2)	RLM (DMA/ R^2)
10	classe 21	18,83/0,953	16,69/0,962	9,19/0,991
	classe 41	18,02/0,964	17,13/0,968	14,60/0,970
	classe 81	11,93/0,986	10,17/0,990	9,20/0,991
20	classe 21	19,62/0,951	17,56/0,957	19,34/0,957
	classe 41	21,10/0,954	17,78/0,966	16,35/0,964
	classe 81	12,16/0,984	10,68/0,990	10,70/0,984
40	classe 21	–	–	–
	classe 41	22,13/0,955	18,23/0,965	25,89/0,876
	classe 81	13,80/0,981	11,85/0,987	12,41/0,980
80	classe 21	–	–	–
	classe 41	–	–	–
	classe 81	16,15/0,973	13,28/0,983	21,62/0,952

A Tabela 4 apresenta os resultados de DMA para os artistas do gênero MPB. Observa-se que o melhor resultado (11,27 de DMA) foi alcançado pela técnica RLM para a classe 81 com a base que possui 10 atributos preditores. Já o pior resultado (DMA de 41,10) também foi obtido pela técnica RLM para a classe 81 com a base que contém 80 atributos preditores. Na maioria dos casos a técnica RLM apresenta os melhores resultados e o k -NN os piores, ficando o RF numa posição intermediária.

Tabela 4: DMA considerando o gênero MPB

Quant. Atributos Preditores	Atributo Classe	k-NN (DMA/ R^2)	RF (DMA/ R^2)	RLM (DMA/ R^2)
10	classe 21	20,66/0,930	17,32/0,956	18,07/0,950
	classe 41	17,90/0,966	18,04/0,966	13,80/0,979
	classe 81	12,06/0,984	11,36/0,986	11,27/0,983
20	classe 21	21,68/0,924	18,23/0,952	21,36/0,941
	classe 41	18,99/0,940	17,46/0,950	12,18/0,985
	classe 81	12,59/0,984	12,42/0,984	12,24/0,985
40	classe 21	–	–	–
	classe 41	17,01/0,970	18,40/0,918	16,66/0,969
	classe 81	19,48/0,955	15,65/0,970	16,37/0,973
80	classe 21	–	–	–
	classe 41	–	–	–
	classe 81	20,37/0,947	17,31/0,958	41,10/0,913

Por fim, a Tabela 5 apresenta os resultados de DMA para os artistas do gênero Rock. Mais uma vez o melhor resultado (DMA de 7,8) foi alcançado pela técnica RLM para a classe 21 para a base com 10 atributos preditores. O pior resultado (DMA de 23,62) foi também alcançado para a RLM, mas para a classe 81 utilizando-se a base com 80 atributos preditores. Considerando-se todos os cenários avaliados, verifica-se que a RLM supera as demais técnicas na maioria dos casos e que o k-NN geralmente apresenta o pior desempenho, ficando novamente o RF com os resultados intermediários.

Tabela 5: DMA considerando o gênero Rock

Quant. Atributos Preditores	Atributo Classe	k-NN (DMA/ R^2)	RF (DMA/ R^2)	RLM (DMA/ R^2)
10	classe 21	10,89/0,987	9,24/0,991	7,80/0,993
	classe 41	16,75/0,963	14,54/0,970	13,00/0,974
	classe 81	14,26/0,968	11,71/0,977	9,07/0,982
20	classe 21	10,52/0,987	9,79/0,989	8,43/0,991
	classe 41	14,59/0,971	14,89/0,972	13,76/0,970
	classe 81	13,36/0,972	11,56/0,976	10,29/0,980
40	classe 21	–	–	–
	classe 41	15,75/0,969	15,49/0,971	16,13/0,950
	classe 81	14,34/0,969	11,64/0,976	13,57/0,971
80	classe 21	–	–	–
	classe 41	–	–	–
	classe 81	16,86/0,957	12,54/0,973	23,62/0,929

Nos diferentes cenários avaliados até aqui para bases formadas por artistas de cada um dos quatro gêneros analisados separadamente (Sertanejo, Pop, Rock e MPB) e para uma base que reúne todos os artistas de todos os gêneros musicais, pode-se verificar que a técnica RLM se destacou em relação às demais, apresentando na maioria das vezes os melhores resultados de DMA e R^2 . Além disso, para todos os casos analisados, o melhor resultado sempre foi obtido a partir da base formada por 10 atributos preditores. É possível observar também que consideráveis variações de desempenho podem ocorrer quando diferentes atributos classe são considerados. No caso dos experimentos aqui realizados, de modo geral, os melhores desempenhos foram alcançados para a classe 81.

Comparando-se os resultados entre os quatro diferentes gêneros (Sertanejo, Pop, Rock e MPB), percebe-se que os

melhores resultados foram encontrados para os artistas do gênero Rock. Esse resultado pode ser explicado pelo fato de esse gênero geralmente ter apresentado as menores variações de posições no *ranking* ao longo do tempo (ver gráfico da Figura 4). Dentre todos os cenários avaliados, a base considerando os artistas de todos os gêneros foi a que alcançou o melhor resultado (7,5 de DMA).

As tabelas apresentadas até o momento mostraram apenas o DMA considerando-se o *ranking* completo em cada um dos cenários avaliados. Para viabilizar uma análise do desempenho dos modelos de regressão para diferentes faixas do *ranking*, gerou-se o gráfico da Figura 5, que apresenta os valores de DMA para todos os cenários em diferentes intervalos de posições: iniciais (entre 1° e 50°), intermediárias (entre 51° e 430°) e finais (entre 431° e 478°). Apenas os resultados de DMA para a RLM são mostrados, já que essa foi a técnica que proveu os melhores resultados. Nesse gráfico, o eixo *x* apresenta cada um dos cenários analisados e o eixo *y* os valores de DMA. Pode-se observar que, exceto para os artistas do gênero MPB no intervalo de 1° e 50°, para todos os demais casos, os resultados de DMA apresentados no gráfico da Figura 5 estão de acordo com as variações de posições no *ranking* apresentadas no gráfico da Figura 4, ou seja, quanto maior a variação de posição, pior o desempenho preditivo dos modelos de regressão e vice-versa.

Vale ressaltar que, no gráfico da Figura 5, o gênero Sertanejo apresentou o valor de DMA igual a 0 no intervalo de posições 431° a 478° porque não houve artistas desse gênero nesse intervalo de posições do *ranking* considerando-se o atributo classe analisado.

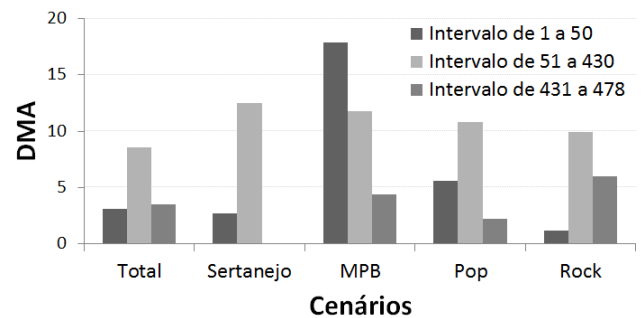


Figura 5: DMA considerando três intervalos, quantidade de atributos preditores de 10 dias e a classe 81.

Na maioria dos cenários os melhores resultados de DMA foram obtidos para o intervalo de posições iniciais (entre 1° e 50°). Ao levar em consideração que os artistas melhores posicionados nos *rankings* são geralmente os mais vistos pela indústria fonográfica, os bons resultados preditivos para esses artistas mostram-se relevantes no auxílio à tomada de decisões nesse meio.

A Tabela 6 apresenta o desempenho da técnica RLM de acordo com a segunda métrica utilizada neste trabalho, o Coeficiente de Determinação R^2 , considerando-se os intervalos de posições iniciais, intermediárias e finais. Pode-se observar que os resultados obtidos para essa métrica estão de acordo com os resultados de DMA apresentados nos gráficos e tabelas anteriores, ou seja, modelos de regressão gerados a partir da RLM se ajustaram muito bem aos dados ($R^2 \geq 0,722$) disponíveis em todos os cenários (exceto

para os artistas do gênero MPB no intervalo de posições iniciais, onde $R^2 = 0,12$, e para os artistas do gênero Sertanejo considerando-se o *ranking* completo, utilizando-se a classe 41 com a base que possui 40 atributos preditores, onde $R^2 = 0,266$ – ver Tabela 2).

Tabela 6: R^2 para os intervalos de posições iniciais, intermediárias e finais

Intervalos	Total	Sertanejo	Pop	Rock	MPB
Intervalo Inicial (1 a 50)	0,838	0,931	0,722	0,995	0,120
Intervalo Intermediário (51 a 430)	0,983	0,974	0,981	0,972	0,973
Intervalo Final (431 a 478)	0,906	–	0,975	0,883	0,886

5. CONCLUSÃO

Os dados contidos nas mídias sociais têm sido utilizados como indicadores da popularidade de artistas e, por isso, vêm sendo utilizados para geração de *rankings*. Um exemplo disso é o *ranking* “Social 50” divulgado na revista Billboard. A geração de *rankings* é importante por permitir aos analistas compreender a dinâmica do meio fonográfico e identificar tendências. Assim, a realização de *rankings* é um problema importante de ser estudado.

Devido à importância da realização de *rankings* no escopo musical, este trabalho avalia técnicas de mineração de dados na predição de *rankings*. Mais especificamente, a partir de dados históricos de posições de artistas em *rankings*, utilizam-se técnicas de regressão para prever as posições de artistas no *ranking* em momentos futuros.

As técnicas de regressão k -NN, RF e RLM foram avaliadas com diferentes parâmetros, número de instâncias e quantidade de atributos preditores. De acordo com as métricas DMA e R^2 , a técnica que gerou os melhores modelos de regressão foi a RLM. Para todos os casos analisados, o melhor resultado sempre foi obtido a partir de uma base de dados formada por 10 atributos preditores. Além disso, consideráveis variações de desempenho foram observadas quando diferentes atributos classe (associados a diferentes datas) foram utilizados. Verificou-se também a variação de desempenho preditivo dos modelos quando os artistas contidos nas bases de dados são separados de acordo com o gênero musical. Dentre os gêneros avaliados, os melhores desempenhos foram obtidos para o Rock, que está entre os gêneros onde os artistas apresentam as menores variabilidades de posição no *ranking* ao longo do tempo.

Por fim, verificou-se que os modelos de regressão fornecem predições próximas aos valores reais, ou seja, com baixas taxas de erro. Por exemplo, considerando a base que contempla artistas de todos os gêneros, o desvio médio absoluto (DMA) foi de 7,5. O bom desempenho dos modelos gerados pode ser confirmado a partir da métrica R^2 , que na maioria dos casos se aproximou de 100%. Considerando os resultados preditivos por intervalos de posições no *ranking* (iniciais – do 1º ao 50º, intermediárias – do 51º ao 430º e finais – do 431º ao 478º), os melhores resultados (1,14 de DMA para a base com artistas do gênero Rock) foram obtidos para os artistas classificados no intervalo de posições iniciais do *ranking*. Esse resultado é interessante pelo fato de a indústria fonográfica geralmente estar mais interessada nos artistas melhores posicionados no *ranking*.

Apesar de outros trabalhos na literatura já terem abordado o problema da predição de *rankings*, os resultados ob-

tidos por eles não são comparáveis aos aqui apresentados porque os dados utilizados para realizar a predição são diferentes daqueles adotados neste trabalho.

Como trabalho futuro sugere-se uma experimentação que considere a sazonalidade específica por datas comemorativas como, por exemplo, natal e carnaval, gerando modelos preditivos a partir de bases de dados históricos que contemplem artistas de gêneros musicais relacionados com essas datas comemorativas.

6. AGRADECIMENTOS

Os autores agradecem à Universidade Federal de Ouro Preto, CAPES, FAPEMIG e CNPq por apoiarem o desenvolvimento desta pesquisa. Agradecem também aos revisores anônimos por seus comentários construtivos, contribuindo para uma versão melhor deste artigo.

7. REFERÊNCIAS

- [1] D. W. Aha. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, 36(2):267–287, 1992.
- [2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [3] N. J. Bryan and G. Wang. Musical influence network analysis and rank of sample-based music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR*, pages 329–334, Miami, Florida, 2011.
- [4] Z.-S. Chen, J.-S. Jang, and C.-H. Lee. A kernel framework for content-based artist recommendation system in music. *IEEE Transactions on Multimedia*, 13(6):1371–1380, 2008.
- [5] J. Grace, D. Gruhl, K. Haas, M. Nagarajan, C. Robson, and N. Sahoo. Artist ranking through analysis of online community comments. Technical report, IBM, 2007.
- [6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.
- [7] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.
- [8] Han, J. and Kamber, M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.
- [9] N. Koenigstein and Y. Shavitt. Song ranking based on piracy in peer-to-peer networks. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR*, pages 633–638, Kobe, Japan, 2009.
- [10] R Development Core Team, R: A Language and Environment for Statistical Computing, Viena, Austria, 2009. [Online]. Available: <http://www.R-project.org/>
- [11] Y. Sun, H. Li, I. G. Councill, J. H. 0002, W.-C. Lee, and C. L. Giles. Personalized ranking for digital libraries based on log analysis. In *Proceedings of the 10th ACM International Workshop on Web Information and Data Management*, pages 133–140, Napa Valley, California, 2008.

- [12] G. L. Toledo and I. I. Ovalle. *Estatística Básica*, page 251. Atlas, São Paulo, 2 edition, 1983.
- [13] Y.-H. Yang, Y.-C. Lin, A. Lee, and H. Chen. Improving musical concept detection by ordinal regression and context fusion. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR*, pages 147–152, Kobe, Japan, 2009.
- [14] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen. A regression approach to music emotion recognition. *IEEE*, 16(2):448–457, 2008.

Recomendação de Conteúdos Multimídia em Ambientes de Convergência Digital: Uma Abordagem Baseada em Mineração de Dados e Web Semântica

Alternative Title: Multimedia Content Recommendation in Digital Convergence Environments: An Approach Based on Data Mining and Semantic Web

Priscilla Kelly M. Vieira
Universidade Federal Rural de
Pernambuco (UFRPE)
Unidade Acadêmica de Garanhuns
Garanhuns, PE, Brasil
priscillakmv@gmail.com

Natasha Queiroz Lino
Universidade Federal da Paraíba
(UFPB)
Centro de Informática
João Pessoa, PB, Brasil
natasha@ci.ufpb.br

RESUMO

Com o advento da TV Digital interativa (TVDi), nota-se o aumento de interatividade no processo de comunicação além do incremento das produções audiovisuais, elevando o número de canais e recursos disponíveis para o usuário. Esta realidade faz da tarefa de encontrar o conteúdo desejado uma ação onerosa e possivelmente ineficaz. A incorporação de sistemas de recomendação no ambiente TVDi emerge como uma possível solução para este problema. Este trabalho tem como objetivo propor uma abordagem híbrida para recomendação de conteúdo em TVDi, baseada em técnicas de Mineração de Dados, integradas a conceitos da Web Semântica, permitindo a estruturação e padronização dos dados e consequente possibilidade do compartilhamento de informações, provendo semântica e raciocínio automático. Para o serviço proposto é considerado o Sistema Brasileiro de TV Digital (SBTVD) e o *middleware* Ginga. Foi desenvolvido um protótipo e realizado experimentos com a base de dados do Netflix, utilizando a métrica de precisão para avaliação. Obteve-se uma precisão média de 30%, utilizando apenas a técnica de mineração. Acoplado-se com as regras semânticas obteve-se precisão média de 35%.

Palavras-Chave

Sistemas de Recomendação, Mineração de Dados, TVDi, Representação de Conhecimento.

ABSTRACT

The emerging scenario of interactive Digital TV (iDTV) is promoting the increase of interactivity in the communication process and also in audiovisual production, thus raising the number of channels and resources available to the user. This reality makes the task of finding the desired content becoming a

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.
Copyright SBC 2015.

costly and possibly ineffective action. The incorporation of recommender systems in the iDTV environment is emerging as a possible solution to this problem. This work aims to propose a hybrid approach to content recommendation in iDTV, based on data mining techniques, integrated to the semantic web concepts, allowing structuring and standardization of data and consequently making possible sharing of information, providing semantics and automated reasoning. For the proposed service it is considered the Brazilian Digital TV System (SBTVD) and the *middleware* Ginga. A prototype has been developed and experiments carried out with a Netflix database. As results, it was obtained an average accuracy of 30% using only the data mining technique. On the other hand, the evaluation including semantic rules obtained an average accuracy of 35%.

Categories and Subject Descriptors H.3.5 [Information Storage and Retrieval]: On-line Information Services – data sharing, web-based services

General Terms

Algorithms, Design.

Keywords

Recommender Systems, *Data Mining*, iDTV, Knowledge Representation.

1. INTRODUÇÃO

Com o advento da TV Digital interativa (TVDi), nota-se o aumento de interatividade no processo de comunicação além do incremento das produções audiovisuais [1], o que torna o ambiente favorável ao desenvolvimento de aplicações dedicadas a esta tecnologia. Neste sentido, a TVDi promoveu e possibilitou um cenário de aumento da quantidade de canais, serviços e conteúdo disponíveis ao usuário. Esta nova realidade faz com que a tarefa de encontrar o conteúdo desejado se torne uma ação onerosa e, em alguns casos, ineficiente. É neste contexto que Sistemas de Recomendação [2; 3; 4] emergem como solução possível para auxiliar esta escolha.

Neste sentido, é proposto neste trabalho um serviço de recomendação de conteúdo no ambiente do Sistema Brasileiro de TV Digital (SBTVD) [5], baseado em Mineração de Dados e em conceitos e técnicas da Web Semântica. Estas duas estratégias foram escolhidas por já serem utilizadas para o propósito deste

trabalho, mas não terem sido avaliadas em conjunto, considerando o ambiente brasileiro [6; 7]. Para a esquematização do serviço de recomendação de conteúdo foi definida e instanciada uma arquitetura, tal como projetada uma ontologia responsável por estruturar dados multimídia minerados.

É importante destacar que as estratégias e arquitetura utilizadas neste trabalho, podem ser aplicadas e avaliadas em diversos outros ambientes com o objetivo de recomendação de conteúdo. Contudo, foi utilizado ambiente de TVDi como forma de demonstração e validação das estratégias propostas.

O restante deste artigo está organizado da seguinte forma: a seção 2 apresenta uma fundamentação teórica de Sistemas de Recomendação de conteúdo. Na seção 3 é explanada a arquitetura do serviço de recomendação, o *Recommender Knowledge TV*, tal como o detalhamento de cada um dos componentes envolvidos. Na seção 4 são destacados os experimentos e discussões. Em seguida, na seção 5, são explanados os trabalhos relacionados. Por fim, a seção 6 traz as conclusões e as perspectivas de trabalhos futuros.

2. SISTEMAS PARA RECOMENDAÇÃO DE CONTEÚDO

Sistemas de Recomendação são comumente aplicados em sistemas web [8], mas nos últimos anos começou-se o estudo de recomendação de conteúdo no ambiente da TVDi por possuir algumas peculiaridades.

A plataforma de TVDi tem um conjunto de requisitos que a torna desafiadora para a construção de sistemas de recomendação de conteúdo. Primeiramente, o ambiente de TV pode ser mono-usuário ou multi-usuário, por isso ao recomendar é difícil identificar o tipo de experiência que está sendo realizada e, consequentemente, identificar o usuário que solicita o serviço. Em segundo lugar, este é um cenário multi-plataforma, onde o conteúdo pode ser entregue a diferentes plataformas computacionais, tais como, set-top-boxes (STBs) com o *middleware* padrão, dispositivos móveis e de computação pessoal em geral. Adicionalmente, o comportamento do usuário ao acessar conteúdos destinados a TVDi é, em geral, passivo, diferindo do comportamento ativo na web. Neste sentido, o serviço de recomendação proposto neste trabalho está de acordo com as características intrínsecas ao ambiente da TVDi.

É importante destacar que Sistemas de Recomendação para a TVDi pode possuir diferentes perspectivas de uso, dentre elas destaca-se: (i) do usuário; para sistemas que objetivam auxiliar o usuário na busca por conteúdos de seu interesse. (ii) do provedor de conteúdo; para auxiliar na construção de grades de programação para maximizar alguma métrica, por exemplo audiência. (iii) de marketing, para dar suporte a veiculação de comerciais para o público alvo correto. Este trabalho foca na perspectiva do usuário.

Adicionalmente, é considerado um ambiente de convergência digital. Uso conjunto de duas plataformas distintas: TVDi e web, fazendo uso de dados da web para compor uma solução na TVDi. Na próxima seção são explanadas algumas características do processo de recomendação de conteúdo proposto neste trabalho.

3. PROCESSO PROPOSTO

Para este trabalho foi considerado o SBTVD [5], incluindo o *middleware* Ginga [9]; camada de software intermediária, presente no receptor, e situada entre as aplicações e o sistema operacional, que oferece uma série de facilidades para o desenvolvimento de conteúdo e aplicativos para TV Digital Interativa [10; 11]. Neste sentido propõe-se uma arquitetura para recomendação de conteúdo em ambientes de convergência (TVDi e web), considerando as duas plataformas em conjunto, a TVDi conectada.

No âmbito deste trabalho é utilizado um processo híbrido de recomendação de conteúdo [12] que consiste no uso conjunto de Filtragem Colaborativa (FC) e Filtragem Baseada em Conteúdo (FBC).

Na estratégia com FC é realizada a troca de experiências entre usuários. São detectados grupos de usuários com comportamentos semelhantes, e a recomendação de conteúdo é realizada entre estes, o que possibilita recomendações inesperadas e novas ao usuário. A FBC parte do princípio de que, se um usuário aprovou um determinado programa, é provável que no futuro goste de outros similares, com as mesmas características [12].

A abordagem híbrida foi escolhida por 3 (três) motivos: (i) validar o objetivo específico de permitir recomendação multimídia em TVDi conectada (ambiente de convergência digital), considerando, assim, os dois cenários definidos na literatura [13], FC e FBC. (ii) Validar a arquitetura genérica de recomendação de conteúdo proposta neste trabalho, garantindo que esta pode ser instanciada para ambos os tipos de recomendação de conteúdo. (iii) Considerar o desafio de recomendação de conteúdo multimídia no ambiente da TVDi, como sendo um ambiente de experiência multiusuário, permitindo a atenuação dos problemas das abordagens em separado.

Considerando tais características explanadas, neste trabalho a tarefa de clusterização é realizada sobre os dados dos históricos dos usuários (FC). Neste contexto, supondo que o usuário y (U_y) solicita uma recomendação, é detectado qual o usuário x (U_x) do cluster possui maior similaridade com o U_y . Um conteúdo visto por U_x é recomendado para U_y .

Como forma de aperfeiçoar a decisão de qual programa recomendar para um U_y , a partir de um U_x , regras semânticas são utilizadas (seção 3.2) e as recomendações são realizadas entre os usuários do mesmo cluster (FBC). As regras detectam qual programa de U_x é mais similar semanticamente com os programas vistos pelo U_y .

Com o objetivo de possibilitar o processo de recomendação descrito, foi especificada e avaliada uma arquitetura para prover o serviço de recomendação de conteúdo, o *Recommender Knowledge TV* (RKTV), introduzido na próxima seção.

3.1 Arquitetura

Sob a perspectiva da execução das operações previstas neste trabalho, tais como a tarefa de Mineração de Dados, considerando o ambiente da TVDi conectada, o *middleware* Ginga e a capacidade de processamento dos decodificadores de sinais, os STBs, este trabalho propõe uma arquitetura cliente-servidor, na qual o módulo cliente coleta dados do usuário,

situados no STB (STB) do usuário e inserido dentro do núcleo comum do *middleware*, e o módulo servidor é responsável pela maior parte do processamento de informações e a disponibilização do serviço. Esta arquitetura foi assim definida por ter sido analisado o funcionamento de aplicações na plataforma da TVDi brasileira, onde não se tem o controle da sobreposição de dados e a seguridade de que o contexto do usuário é armazenado ao ser alterado o canal. Uma forma de garantir a manutenção do contexto do usuário é inserindo esta funcionalidade no Núcleo Comum do Ginga [14].

Na Figura 1 são apresentados os componentes de cada um dos módulos da arquitetura RKTv (*Recommender Knowledge TV*).

O módulo cliente coleta dados por meio de dois componentes (Figura 1): (i) *Monitor Agent*: Monitora o comportamento dos usuários do STB em relação ao consumo do conteúdo exibido na TVDi conectada. (ii) *Provider Agent*: Tem a função de capturar as informações sobre a programação visualizada pelo o usuário.

O *Monitor Agent* [14], para monitorar o comportamento dos usuários do STB, está em comunicação constante com o componente *Tuner* [9] do núcleo do *middleware* Ginga (*Ginga Common Core*), responsável pela sintonização dos canais.

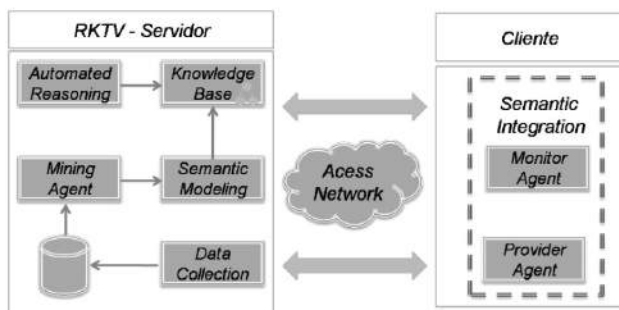


Figura 1. Arquitetura Cliente-servidor do RKTv.

O módulo *Provider Agent* [14] recebe as interações do usuário por meio do módulo *Monitor Agent* e captura os metadados da programação via web ou componente SI do Ginga Common Core [9], responsável por extrair os metadados transmitidos pelos provedores de conteúdos.

Para o objetivo principal deste trabalho alguns dados mínimos devem ser coletados a partir do *Provider Agent*. Estes devem ser utilizados nos demais módulos especificados, auxiliando o serviço de recomendação de conteúdo. Para isto, foram adaptados os metadados definidos em [15], reduzindo seu número e adequando-os ao contexto deste trabalho. Na Tabela 1 são listados os metadados a serem obtidas e suas respectivas funções.

O módulo servidor é composto por cinco módulos:

- *Data Collection*: Captura os dados da programação das emissoras e dos usuários (a partir dos componentes inseridos e localizados no núcleo do Ginga Common Core);
- *Mining Agent*: Responsável por todo o processo de Mineração de Dados sobre os dados coletados do STB e da web sobre o usuário e provedores de conteúdo;
- *Semantic Modeling*: Torna homogeneia a descrição do conteúdo multimídia minerado, estruturado por meio de ontologias;
- *Knowledge Base*: Armazena todo o conhecimento gerado no processo proposto em forma de instâncias de ontologias;

- *Automated Reasoning*: Responsável por inferir as informações presentes nas bases de conhecimento, gerando recomendações de conteúdo compatíveis aos interesses do usuário.

O Módulo *Data Collection* tem como função principal capturar dados: conteúdo visualizado pelo usuário. Estes dados são baseados nos metadados transmitidos junto ao conteúdo audiovisual, por meio do componente *Service Information* (SI) do *middleware* Ginga [14], tal como da web. Os dados provenientes do histórico do usuário são enviados pelo componente *Semantic Integration* (Figura 1), via canal de retorno.

Tabela 1. Metadados Extraídos a partir do *Monitor Agent*

Dado	Função
STB_id	Identificar o STB
Program_name	Identificar nome do programa
Channel	Identificar o canal da TV
Program_genre	Gênero do programa
Start_time	Data e horário em que o usuário começou a assistir ao programa
End_time	Data e horário em que o usuário terminou de assistir ao programa

Os metadados capturados a partir do módulo *Data Collection* (Listados na Tabela 1) são depositados em um banco de dados que é submetido ao processo de Mineração de Dados. O Módulo *Mining Agent* é responsável por aplicar a tarefa de Mineração de Dados, tal como a de pré-processamento. Neste trabalho instanciamos apenas a tarefa de clusterização [16] de dados, mas qualquer tarefa de mineração pode ser instanciada nesta arquitetura. Na próxima subseção será detalhada a metodologia utilizada para a criação da ontologia para estruturar conteúdos multimídia minerados, permitindo o compartilhamento de informações tal como raciocínio semântico.

3.2 Representação Semântica

A Web Semântica pode ser definida como uma extensão da web sintática (web simplificada, sem semântica para os dados), na qual é dada à informação um significado [17], permitindo: (i) processamento automático por máquinas, por meio da padronização da representação do conhecimento, (ii) integração de dados, (iii) reuso de dados e (iv) inferência de conhecimento.

A proposta da Web Semântica é baseada nos conceitos de Representação do Conhecimento (RC), subárea da Inteligência Artificial que avalia como o conhecimento pode ser representado simbolicamente e manipulado de forma automática por máquinas [18].

Formalizar o conhecimento permite a interoperabilidade de dados, assim como o raciocínio automático por máquinas, podendo agregar semântica a dados brutos. Para esta formalização, pode-se fazer uso de ontologias [19; 20].

Neste trabalho é utilizada a abordagem da Mineração de Dados para as ontologias, onde o conhecimento proveniente do processo da tarefa de clusterização é representado por meio de uma ontologia, descrita em OWL, chamada OntoRKTv (Onto –

Ontologia e RKTV – *Recommender Knowledge TV*), possibilitando assim, o raciocínio automático sobre os dados e o compartilhamento de dados.

Neste sentido, pode-se realizar raciocínio, por exemplo, sobre gêneros de programas, recomendando, portanto, programas com gênero equivalente ao que o usuário costuma visualizar ou programas com gênero em um subdomínio que o usuário costuma visualizar.

Dentre a gama de metodologias propostas para a construção de ontologias [21; 22], neste trabalho utilizou-se a metodologia 101 especificada por Noy e McGuinness [23] por ser uma metodologia já estabelecida e ter sua implementação e documentação baseada na ferramenta Protegé [24], amplamente difundida para a construção de ontologias.

Seguindo os passos definidos em [23] foi gerada a ontologia da Figura 2, com destaque para o nome das classes e os relacionamentos. As propriedades foram suprimidas para facilitar a visualização geral da OntoRKTV.

Dentre as classes modeladas tem-se: (i) Cluster: Agrupamento de STB com uso semelhante. (ii) STB: Um membro que pode ser pertencente a um cluster. (iii) History: Os membros dos clusters possuem históricos de uso; (iv) PeriodOfDay: Os membros de um cluster são definidos pelo padrão de uso por turno.

As classes ContentDescription e TemporalUnit foram definidas em uma outra ontologia [15] que descreve todas os conceitos importantes para descrever conteúdos multimídia. As classes descritas são utilizadas como links entre os conceitos de clusterização e de dados multimídia para recomendação, indicando, assim, o reuso de ontologias para compor a ontologia descrita neste trabalho. Neste sentido, ContentDescription, ou descrição de conteúdo, é a classe que representa um ponto chave agregador de diversas informações descritivas acerca do conteúdo transmitido. A classe TemporalUnit detalha questões de tempo sobre a programação.

Algumas regras semânticas foram definidas e desenvolvidas sobre a OntoRKTV, para apoiar o processo de recomendação no ambiente da TVDi:

I.isPartOfCluster: dois STBs U_1 e U_2 pertencem a um mesmo cluster se: (i) U_1 visualiza conteúdo semelhante ao conteúdo visualizado por U_2 .

II.neighbor: U_1 é vizinho com forte interseção com U_2 se: (i) O vizinho com maior número de conteúdos similares à U_2 é U_1 ; e (ii) U_1 pertence ao mesmo cluster que U_2 .

A regra *isPartOfCluster* é uma consequência da tarefa de Mineração de Dados. Dentro de cada cluster, podem-se utilizar as regras que se baseiam diretamente nos relacionamentos oriundos da definição de ontologias:

III. *isEquivalentTo*: uma instância D_1 é equivalente a uma instância D_2 se (i) D_1 é equivalente a um conceito k na ontologia OntoRKTV; (ii) D_2 é equivalente a um conceito w na mesma ontologia; e (iii) k e w estão ligados por um relacionamento *equivalentClass* na OntoRKTV.

IV. *isSubConceptOf*: um dado D_1 é uma especialização de um dado D_2 se (i) D_1 é idêntico a um conceito k na ontologia OntoRKTV; (ii) D_2 é idêntico a um conceito w na mesma ontologia; e (iii) k e w estão ligados por um relacionamento *subClassOf* na ontologia OntoRKTV.

V. *isSuperConceptOf*: um dado D_1 é uma generalização de um dado D_2 se: (i) D_1 é idêntico a um conceito k na ontologia OntoRKTV; (ii) D_2 é idêntico a um conceito w na mesma ontologia; e (iii) k e w estão ligados por um relacionamento *superClassOf*.

VI. *isInstanceCloseTo*: Dois dados D_1 e D_2 são instâncias próximas se: (i) D_1 é idêntico a uma instância k na ontologia OntoRKTV e D_2 é idêntico a uma instância w na mesma ontologia; e (ii) k e w são instâncias do mesmo conceito. No contexto deste trabalho, instância é o valor atribuído a uma classe ou propriedade de uma ontologia.

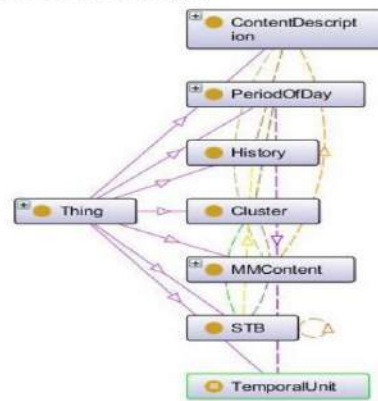


Figura 2. Ontologia OntoRKTV.

As regras definidas em III –VI são utilizadas para avaliação semântica do conteúdo a ser recomendado de um usuário para outro. Neste sentido, avalia-se se dois conteúdos possuem proximidades semânticas. Por exemplo, se um filme é de um subgênero do gênero que um usuário costuma visualizar.

Na próxima subseção será detalhado o módulo *Mining Agent* da arquitetura RKTV.

3.3 Mining Agent

O módulo *Mining Agent* (Figura 1) é responsável por todo o processo relacionado com a descoberta de conhecimento, com foco principal em pré-processamento e Mineração de Dados [25].

Tarefas de Mineração de Dados [26] classificam o padrão que se deseja obter da base de dados. Podem ser classificadas como: (i) Regras de Associação, (ii) Clusterização, (iii) Sumarização, (iv) Classificação e (v) Regressão, dentre outros. Neste trabalho damos foco à tarefa de Clusterização [26; 27], com o objetivo de avaliar o comportamento da recomendação no ambiente de convergência em análise, dado que um outro trabalho [28] avalia este comportamento por meio da tarefa de Regras de Associação. Neste sentido não é objetivo deste trabalho propor algoritmos ótimos de clusterização, mas mostrar a viabilidade da aplicação desta tarefa no ambiente da TVDi.

Adicionalmente, clusterização é utilizada como parte do processo de recomendação proposto por ser amplamente utilizada em sistemas de recomendação e ser considerada de boa precisão para o contexto de recomendação [29].

Existem diversos algoritmos na literatura com foco em clusterização de dados [30; 31]. Neste trabalho é dada ênfase à técnica k-means [30; 31; 32]. Este foi escolhido por ser um algoritmo popular, simples, direto e com custo linear [30].

Para realizar a clusterização de instâncias de objetos é necessário o uso de funções que calculam a similaridade entre os objetos: No contexto de clusterização existem diversas funções que objetivam calcular a distância entre dois objetos [33], dentre estas: euclidiana, *levenshtein*, *hamming*, cosseno. Para a escolha da função é necessário avaliar os dados com relação as características, tipos de dados que serão comparados e os objetivos da comparação [33].

Neste sentido, foi avaliado o comportamento da função de distância euclidiana. Esta foi utilizada nos experimentos explanados na próxima seção (seção 4).

A distância euclidiana é expressa segundo a Equação 1. Nesta, x e y são vetores com n atributos numéricos. Assim $(x_a - y_a)$ é a diferença entre os atributos de x e y na posição a . A distância euclidiana é a raiz quadrada da soma do quadrado da diferença $(x_a - y_a)$, para todo a .

Considerando o contexto deste trabalho, os vetores, x e y , que são parâmetros para a função da Equação 1, representam o histórico do usuário x e do usuário y . Na seção 4.1 é destacado o processo para a geração destes vetores. Na próxima seção são detalhados os experimentos para a validação deste trabalho.

Equação 1. Distância euclidiana

$$E(x, y) = \sqrt{\sum_{a=1}^n (x_a - y_a)^2}$$

4. Experimentos

A arquitetura RKTv (Figura 1) é genérica, podendo ser instanciada por diversas técnicas e com o uso de diferentes tecnologias. Como forma de demonstrar a viabilidade da arquitetura, foi realizada uma implementação de cada um dos módulos especificados no componente servidor. Neste sentido, foi desenvolvido um protótipo passível de recomendação de conteúdo.

O foco deste trabalho não foi a especificação/implementação do componente do núcleo comum do Giga (Figura 1). Este foi especificado e implementado em outro trabalho [14], tal qual as necessidades do RKTv, definidas neste trabalho, bem como do projeto mais abrangente do qual o RKTv faz parte.

O módulo *Mining Agent* foi desenvolvido em Matlab, desde o pré-processamento dos dados até a fase da tarefa de Mineração de Dados (onde foi utilizado o algoritmo k-means para clusterização).

As funcionalidades do módulo *Semantic Modeling* foram desenvolvidas por meio da ferramenta Protegé [24]. Para raciocínio automático utilizou-se a API Jena [34].

Por fim, foi desenvolvido em Java um protótipo baseado no RKTv, instanciando e integrando os seus módulos, implementando o processo de recomendação proposto neste trabalho.

Com o objetivo de validar a aplicabilidade do processo definido para o RKTv da plataforma KTV para TVDi, foram realizados experimentos para dar subsídios que, de forma prática, demonstrassem a aplicabilidade e viabilidade deste trabalho.

Nesse sentido foi utilizada uma base de dados do serviço de stream de vídeos Netflix [35] com o registro de histórico de usuários. A utilização da base de dados disponibilizada pelo Netflix se dá devido esta fazer parte de um subdomínio da TVDi, no caso, filmes. Estas características agregam valor a esta pesquisa e experimentos feitos, principalmente por ser uma base de dados real.

A base original de dados do Netflix disponibiliza as seguintes informações: IDFilme (identificação do filme), CustomerID (identificação do usuário), Rating (Número de 'estrelas' atribuídos a um filme por um cliente), Título do filme, Ano que um filme foi lançado e Data de uma classificação.

Os experimentos foram utilizados com a base original (com os atributos listados), e uma repetição do experimento com a base alterada por meio de um enriquecimento semântico. Este enriquecimento se deu por informações retiradas do DBPedia [36], por meio de consultas em SPARQL [37]. Os atributos coletados foram: (i) Diretor, nome do diretor responsável pelo filme. (ii) Ator, atores que participaram do filme; (ii) Categoria, categoria na qual o filme se enquadra.

4.1 Pré-processamento dos dados

No processo de descoberta de conhecimento, uma das etapas que antecede a Mineração de Dados é o pré-processamento. Este precisa ser utilizado para que os dados sejam preparados de acordo com a técnica a ser utilizada. Para a realização da clusterização de usuários, foi gerada uma matriz com uma coluna para o código do usuário e colunas para os filmes vistos por cada usuário. A Tabela 2 demonstra esta transformação. Cada linha da tabela representa um vetor de interesses de um usuário, sendo assim, dois usuários com interesses semelhantes possuem vetores próximos. Desta forma, na Tabela 2, tem-se que o usuário de identificação "69867", assistiu ao filme "Congo", uma vez, e aos filmes "Mississippi Burning" e "The Santa Clause", zero vezes.

Tabela 2. Exemplo de vetores representativos dos interesses dos usuários

user_id	"Congo"	"Mississippi burning"	"the santa clause"
69867	1	0	0
437	1	1	0

Com o objetivo de realizar recomendações, a base foi fatiada por períodos, por meio do atributo "Data de Classificação". Isto foi realizado para que fosse possível realizar recomendações futuras, podendo confirmar quais delas o usuário realmente aceitaria. Neste caso, foram realizados 30 fatiamentos aleatórios, de seis meses cada.

Considerando uma das fatias, com os 6 primeiros meses de uso do sistema ("08-12-1999" - "08-05-2000"). Após o pré-processamento temporal, os dados foram sumarizados da seguinte forma: 23136 Avaliações, 3432 Usuários, 187 Filmes, 168 Categorias, 168 Atores e 155 Diretores.

Analisando os dados, observou-se a ocorrência de *sparsity problem* [8], bastante conhecido em abordagens colaborativas, que caracteriza-se pela dispersão dos itens avaliados pelo o indivíduo em relação a todos os itens do sistema. Este problema ocorre em situações onde o número de itens é muito grande

tornando impraticável as suas avaliações. Este problema se deu pela grande variedade de filmes a serem avaliados, gerando uma matriz esparsa de dados (muitas dimensões com poucas avaliações).

Para comprovar esta hipótese o algoritmo k-means foi executado com todas as dimensões geradas, com k, número de clusters, variando de 2 – 400 e o máximo de média de *Silhouette Coefficient* (Coeficiente de Silhueta) [27], utilizado para avaliação de clusters, foi 0.05, quando o indicado é superior a 0,5 [25, 38].

Portanto, foi realizada a redução da variabilidade de filmes, selecionando apenas aqueles que foram visualizados por pelo menos 10% dos usuários do sistema. Como exemplo de redução de uma das execuções do experimento tem-se: 18 dimensões de filmes, 17 de atores, 17 de diretores e 17 categorias.

Este pré-processamento foi realizado em todas as 30 partições avaliadas no processo de validação deste trabalho. Na próxima subseção será avaliado o processo de recomendação proposto neste trabalho.

4.2 Execução do Processo Proposto do Mining Agent

Após o pré-processamento dos dados foram aplicadas as técnicas para a detecção do agrupamento de dados. O algoritmo foi executado considerando o histórico de cada usuário.

Para o processamento do algoritmo k-means, o k variou de 2 – 400. Para diminuir a distorção causada pela obtenção de mínimos locais, influência da escolha dos centroides iniciais, [27], foram realizados testes com 5 (cinco) sementes aleatórias diferentes para cada valor de k. Após este processo, selecionou-se a semente que trouxe melhor qualidade aos clusters. Esta qualidade foi avaliada de acordo com duas funções: (i) Média do coeficiente de silhueta [27], que avalia combinação entre coesão e separação dos grupos, (ii) Média dos erros quadráticos [27], que avalia a coesão dos agrupamentos gerados. A média do coeficiente de silhueta foi avaliado segundo [27, 38].

A partir de k = 157, as alterações na média do coeficiente de silhueta foram pouco significativas, com crescimento lento, como demonstrado no gráfico da Figura 5. Neste sentido, os experimentos foram realizados com k=157.

Além do coeficiente de silhueta, foi avaliada a medida de erro quadrático, que tende a zero, quando k tende a infinito [26]. É importante destacar que a decisão do número de grupos é primordial. Não se deve quebrar um cluster em dois distintos, nem tão pouco unir clusters que essencialmente são distintos [16]. Desta forma, avalia-se o erro quadrático onde ocorre um pequeno “joelho” e uma diminuição na velocidade que o erro decreta [27], como destacado no gráfico da Figura 6.

É importante destacar, que neste trabalho foram avaliadas duas medidas para seleção do k ideal, no entanto, estas são independentes, convergindo para um mesmo resultado aproximado, k = 157. Contudo, outros valores de k, aquém de 157, podem ser igualmente escolhidos, no entanto, observou-se um ganho baixo, dado que o número de clusters unitários torna-se bastante presente e pouco representativa para o contexto de recomendação.

Na execução do experimento com a base de dados enriquecida, obteve-se resultados semelhantes, não houveram melhoras

significativas com relação ao coeficiente de silhueta, tão pouco uma redução na soma dos erros quadráticos. Isto se deu devido a características dos dados enriquecidos, pois poucos filmes possuem mesmo ator, diretor ou categoria. O enriquecimento realizado por meio do DBPedia, foi bastante específico, o que não influenciou positivamente no agrupamento dos dados. Esta característica do enriquecimento resultou em uma matriz esparsa, mesmo depois do pré-processamento explanado na subseção 4.1. Neste sentido, os dados provenientes do enriquecimento foram melhor utilizados com as regras semânticas, visto na seção 4.3.

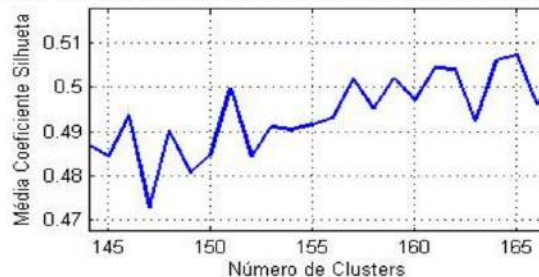


Figura 5. Gráfico Média Coeficiente Silhueta.

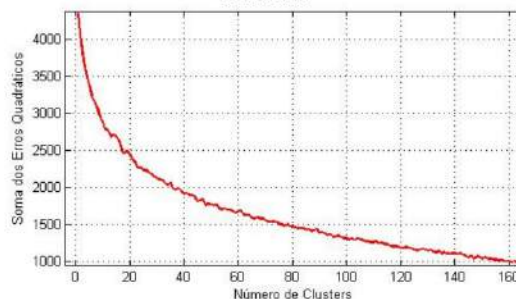


Figura 6. Gráfico Soma dos Erros Quadráticos

4.3 Uso das Regras Semânticas

Após o processo de clusterização (agrupamento sintático), foi integrado ao processo de recomendação do conteúdo o uso das regras semânticas definidas e formalizadas na seção 3.2. Estas regras foram implementadas e avaliadas neste trabalho.

No contexto da base de dados avaliada neste trabalho, as regras definidas em I - VI (seção 3.2) foram aplicadas sobre os atributos que representam a categoria dos filmes, atores e diretores. Considerando os dados da Tabela 3, por exemplo, não ocorre especialização/generalização de gêneros, mas ocorre equivalência de diretores entre os filmes “Yojimbo” e “Seven Samurai” e entre “Presumed Innocent” e “The Devil’s Own”.

No contexto deste trabalho as regras semânticas foram utilizadas com o objetivo de reduzir o grupo de filmes a serem recomendados, tal como aumentar a precisão do sistema, como descrito e avaliado na próxima subseção.

4.4 Análise

Para analisar a recomendação do processo proposto neste trabalho, foi utilizada a métrica *Precision* (Precisão) [39], medindo a porcentagem de itens recomendados de forma correta ao usuário. No contexto deste trabalho, foi realizada uma média da precisão da recomendação realizada aos usuários que foram selecionados aleatoriamente durante os experimentos.

Neste sentido, a precisão das recomendações para cada usuário, segue a expressão definida na Equação 2. A Precisão (P) por usuário é a divisão entre as o número de recomendações que o sistema acertou ($R_{Acertadas}$), pelo número de recomendações realizadas pelo processo (R_{Total}).

Tabela 3. Exemplo do resultado de regras semânticas aplicadas durante os experimentos

Nome do filme	Categoria	Diretor
Yojimbo	Jidaigeki	Akira Kurosawa
Seven Samurai	Fictional samurai	Akira Kurosawa
Presumed Innocent	Works by Scott Turow	Alan J. Pakula
The Devil's Own	Troubles	Alan J. Pakula

Equação 2. Expressão de precisão por usuário

$$p = \frac{R_{Acertadas}}{R_{Total}}$$

A precisão total do processo proposto está definida na expressão da Equação 3. Precisão total do processo proposto neste trabalho (PT) é a média aritmética das precisões de cada usuário, tal que n é o número de usuários.

Equação 3. Expressão da precisão total

$$P_t = \frac{\sum_{i=1}^n \frac{R_{Acertados i}}{R_{Totais i}}}{n}$$

Os acertos foram avaliados a partir do histórico de cada usuário, e das informações provenientes da própria base do Netflix. O processo descrito na subseção anterior foi replicado 30 vezes, com períodos diferentes. Em média, obteve-se uma precisão de 30,04%, sem considerar o uso das regras semânticas.

Adicionando a análise semântica, como previsto no processo de recomendação de conteúdo proposto neste trabalho, obteve-se um refinamento do conjunto de filmes recomendados pelo RKTv, atingindo, em média, 35,12% na métrica de precisão. Este refinamento se dá pela diminuição do conjunto de elementos a serem recomendados. É importante destacar que o pequeno ganho de precisão da recomendação apenas com a clusterização e da recomendação seguida da avaliação semântica se deu pelas características do enriquecimento semântico dos dados. Estes possuem poucas interseções de categorias, tal como poucos mapeamentos exatos definidos nas instâncias na OntoRKTv. Na próxima seção são destacados trabalhos que possuem similaridades com a pesquisa explanada neste artigo.

5. TRABALHOS RELACIONADOS

Aroyo [6] propõe a incorporação da semântica na TVDi, considerando-a no ambiente da web, a fim de possibilitar o desenvolvimento de aplicações distribuídas e a personalização do serviço. Aroyo baseia seu trabalho de personalização nos conceitos da Web Semântica e de Representação do Conhecimento. Assim como Aroyo, nossa proposta também é baseada nos conceitos da Web Semântica, no entanto, propomos recomendação no ambiente do SBTVD, incorporando no *middleware* Ginga um módulo que dá suporte ao processo de recomendação proposto.

Ávila [7] apresenta um sistema integrado ao *middleware* Ginga baseado em mineração por regras de associação, entretanto sem nenhuma estruturação semântica do conteúdo. Diferentemente, este trabalho propõe o uso de ontologias para estruturar o conhecimento após o processo de clusterização dos dados.

Neto[40] e Kim [29] também propõem recomendação de conteúdo utilizando Mineração de Dados. O sistema de recomendação proposto por Kim, Recommendation System of IPTV TV Program Using Ontology and K-means Clustering, apresenta uma solução baseada em mineração de dados por clusterização, mas neste caso a mineração ocorre sobre ontologias.

6. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou uma proposta de abordagem baseada em conhecimento e Mineração de Dados para recomendação de conteúdo em TV Digital interativa conectada. Utilizando o *middleware* Ginga como estudo de caso, propôs-se a especificação de um processo de recomendação que integra conceitos e técnicas da Web Semântica e Mineração de Dados para filtragem híbrida de conteúdo. É importante destacar que inúmeras abordagens e técnicas de recomendação de conteúdo podem ser expressas por meio da arquitetura do RKTv. Neste trabalho, é iniciada a investigação sobre o comportamento de técnicas de clusterização e semântica em conjunto.

Como trabalho futuro, é importante destacar a necessidade de avaliação de outras tarefas de Mineração de Dados, tal como a extensão da OntoRKTv. Adicionalmente, destaca-se a necessidade de especificação do raciocínio automático a ser realizado sobre a ontologia. Por fim, ainda são necessários experimentos para ratificar a importância da OntoRKTv no processo de recomendação. Adicionalmente, é possível incorporar ao processo de recomendação deste trabalho a análise sobre a audiência dos conteúdos, como por exemplo em [41].

7. REFERÊNCIAS

- [1] Médola, A. S. L.(2009) Televisão Digital Brasileira e os Novos Processos de Produção de Conteúdos: Os Desafios para o Comunicador. Revista da Associação Nacional dos Programas de Pós-Graduação em Comunicação | E-compós, Brasília, v.12, n.3, set./dez.
- [2] Resnick, P. and Varian, H. R. (1997) Recommender systems. Communications of the ACM. ACM 40, 3 (Mar. 1997), 56-58.
- [3] Xu, M.; Berkovsky, S.; Ardon, S.; Triukose, S.; Mahanti, A.; Koprinska, I. Catch-up TV Recommendations: Show Old Favourites and Find New Ones. RecSys '13. p. 285-294.

- [4] Azaria, A.; Hassidim, A.; Kraus, S.; Eshkol, A.; Weintraub, O.; Netanel, I. Movie Recommender System for Profit Maximization. *RecSys '13*, p. 12-128.
- [5] SBTVD – Sistema Brasileiro de TV Digital. Ministério das Comunicações. Disponível em: <<http://sbtvd.cpqd.com.br/>>. Acessado em Janeiro de 2014.
- [6] Aroyo, L.; Bellekens, P.; Björkman, M.; Houben, G.; Akkermans, P.; Kaptein, A.; (2007) SenSee Framework for Personalized Access to TV Content. *European Conference on Interactive TV (EuroITV 2007)*, 156-165.
- [7] Ávila, P. M. (2010) RecommenderTV: Suporte ao Desenvolvimento de Aplicações de Recomendação para o Sistema Brasileiro de TV Digital. Dissertação de Mestrado. Universidade Federal de São Carlos, Brasil.
- [8] Torres, R. Personalização na Internet. Editora Novatec, 2004.
- [9] Souza Filho, G. L. de, Leite, L. E. C., Batista, C. E. C. F. Ginga-J: The Procedural Middleware for the Brazilian Digital TV System. In: *Journal of the Brazilian Computer Society*, no 4, Vol 12, (ISSN 0104-6500) pp. 47-56. Março, 2007. Porto Alegre, RS, Brasil.
- [10] Soares, L. F.; Lemos, G. (2007) Interactive Television in Brazil: System Software and the Digital Divide. In *European Interactive TV Conference - EuroITV2007*. Amsterdam.
- [11] Lekakos, G.; Chorianopoulos, K.; Doukidis, G. *Interactive Digital Television: Technologies and Applications*. IGI Publishing, 2007.
- [12] Chorianopoulos, K.. Personalized and mobile digital TV applications. *Proc. 2008 Multimedia Tools and Applications*, Volume 36, Issue 1-2. Kluwer Academic Publishers, pp. 1-10, 2008.
- [13] Adomavicius, G.; Tuzhilin, A. Towards the Next Generation of Recommender Systems: A survey of the State-of-the-Art and Possible Extensions. *Knowledge and Data Engineering, IEEE Transactions*, Volume 17, Issue 6, page(s): 734– 749, June 2005.
- [14] Silva, C. F.; Vieira, P. K. M.; Lino, N. C. Q. Semantic Integration - Uma Extensão do Núcleo do Middleware Ginga. SBSI, 2013.
- [15] Araújo, J. P. C. (2011) CoreKTV - Uma infraestrutura baseada em conhecimento para TV Digital Interativa: um estudo de caso para o middleware Ginga. Dissertação de Mestrado. Universidade Federal da Paraíba.
- [16] Berkhin, P.. Survey of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, CA, 2002.
- [17] Berners-Lee, T.; Lassila, O.; Hendler, J. The semantic web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 2001.
- [18] Russell, S. and Norvig, P. *Artificial Intelligence - A Modern Approach*. 2 edição, Prentice Hall, 2002.
- [19] Breitman, K.. *Web Semântica: O Futuro da Internet*. 1. ed. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora S.A., 2005. v. 1. 190 p.
- [20] Guarino, N. (1995) Formal Ontology, Conceptual Analysis and Knowledge Representation, *International Journal of Human-Computer Studies*, 43(5-6):625–640.
- [21] Fernández-López, M.; Gómez-Pérez, A. Overview and analysis of methodologies for building ontologies. In: *Journal The knowledge Engineering Review*. Vol 17, pp 129-156. New York, 2002.
- [22] Gómez-Pérez, A. Evaluation of taxonomic knowledge in ontologies and knowledge bases. In *Proc. Of the 12th Workshop on Knowledge Acquisition, Modeling and Management, KAW'99*. Voyager Inn, Banff, Alberta, Canada. 1999.
- [23] Noy, F. N.; McGuinness, D. L. *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.
- [24] Protegé. Disponível em: <<http://protege.stanford.edu/>>. Acessado em Junho de 2014.
- [25] Kononenko, I. & Kukar, M. *Machine learning and data mining*. Chichester, UK: Horwood Publishing, 2007.
- [26] Han, J. and Kamber, M. *Data Mining Concepts and Techniques*. 2a Edição, Editora Elsevier, Reino Unido. 2006.
- [27] Tan, P.N., Steinbach, M., Kumar, V. *Introduction to Data Mining*, Addison-Wesley, 2006.
- [28] Patrício Júnior, J. C. A. Mining Knowledge TV: Uma Abordagem de Ambiente de KDD com Ênfase em Mineração de Dados no Ambiente da Knowledge TV. Dissertação de Mestrado. Universidade Federal da Paraíba, 2012.
- [29] Kim, J., Kwon, E., Cho, Y., Kang, S. Recommendation System of IPTV TV Program Using Ontology and K-means Clustering. *Second International Conference, UCMA 2011, Daejeon, Korea, April 13-15, 2011. Proceedings, Part II*.
- [30] Berry, M.J.A.; Linoff, G. *Data mining techniques*. John Wiley & Sons, Inc. 1997.
- [31] Ochi, L. S., Dias, C. R., Stênio, S. F.. Clusterização em Mineração de Dados. *Escola Regional de Informática Rio de Janeiro. Espírito Santo. Mini Curso. Rio das Ostras*, 2004.
- [32] Jain, A. K., Murty, M. N., and Flynn, P. J.. Data clustering: a review. *ACM Comput. Surv.*, 1999. 31(3):264–323.
- [33] Wang, H.; Nearest Neighbor by Neighborhood Counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no.6, pp. 942-953. 2006.
- [34] Jena. Disponível em: < <http://jena.sourceforge.net/>>. Acessado em Janeiro de 2014.
- [35] Netflix. Disponível em <www.netflix.com>. Acessado em Junho de 2014.
- [36] BDPedia. Disponível em <<http://pt.dbpedia.org/>>. Acessado em Janeiro de 2014.
- [37] DuCharme, B. *Learning SPARQL*. O'Reilly Media, 2011.
- [38] Khalilian, M., Mustapha, N., Suliman, N. , Mamat, A. A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets. *IMECS*, 2010. Hong Kong.
- [39] Montaner, M. *Collaborative Recommender Agents Based on Case-Based Reasoning and Trust*. 2003.
- [40] Neto, M. M., Cardoso, D., Teixeira, C. T., Cortés, M. Abordagem Combinada para Recomendação Personalizada Utilizando o Guia de Programação Eletrônico. *CSBC' 2010*, Belo Horizonte, MG, 2010.
- [41] Basilio, A. C A.; Moreno, M. F.; Barrere, E. Análise de Interação e Audiência em Sistemas de TV Digital Interativa. *Webmedia*, 2012.

Um Modelo de Recomendação de Arquivos para Sistemas de Armazenamento em Nuvem

Alternative title: A File Recommendation Model For Cloud Storage Systems

Ricardo Batista
Rodrigues
Centro de Informática
Universidade Federal de
Pernambuco (UFPE)
Recife, PE, Brasil
rbr@cin.ufpe.br

Rodrigo E. Assad
Universidade Federal Rural de
Pernambuco (UFRPE)
Recife, PE, Brasil
assad@usto.re

Carlo M. R. da Silva
Centro de Informática
Universidade Federal de
Pernambuco (UFPE)
Recife, PE, Brasil
cmrs@cin.ufpe.br

Vinicius C. Garcia
Centro de Informática
Universidade Federal de
Pernambuco (UFPE)
Recife, PE, Brasil
vcg@cin.ufpe.br

Frederico A. Durão
Universidade federal da Bahia
(UFBA)
Salvador, BA, Brasil
fred@gmail.com

Silvio R. L. Meira
Centro de Informática
Universidade Federal de
Pernambuco (UFPE)
Recife, PE, Brasil
srlm@cin.ufpe.br

RESUMO

O desenvolvimento tecnológico vivenciado nos últimos anos proporcionou o crescimento do universo digital de forma exponencial, parte desse universo digital encontra-se armazenado em sistemas de armazenamento em nuvem. A cada dia surgem mais destes sistemas, que oferecem o armazenamento de dados de forma distribuída com alta taxa de disponibilidade, o que tem impulsionado cada vez mais usuários a migrarem seus dados para a nuvem. No entanto, a grande quantidade de arquivos armazenada nestes sistemas dificulta a filtragem de conteúdo relevante, demandando tempo e trabalho por parte do usuário na busca por arquivos com conteúdo similar às suas preferências. Diante deste cenário, esta pesquisa propõe um modelo de recomendação para sistemas de armazenamento em nuvem, que tem como objetivo utilizar características da nuvem associadas à técnica de recomendação baseada em conteúdo.

Palavras-Chave

Recomendação, armazenamento em nuvem, computação em nuvem.

ABSTRACT

The technological development in recent years has experienced the exponentially growth of the digital universe, part of this digital universe lies stored in cloud storage systems. With each day, more of these systems come out, offering data

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

storage in a distributed manner with the proposal to provide high availability rate, what has driven more and more users who have migrated your data to the cloud. However, the large amount of files stored in these systems makes it difficult to filter relevant content, requiring time and labor by the user in searching for files with similar content to your preferences. Face of this scenario, this study proposes a model for recommendation of files in cloud storage systems, which aims to use cloud features associated with the technique of content-based recommendation.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Systems; H.3 [Information Storage and Retrieval]: Miscellaneous

General Terms

Theory

Keywords

Recommendation, cloud storage, cloud computing.

1. INTRODUÇÃO

Vivemos em uma era de efervescência informacional, a cada dia se produz mais informação e, geralmente, estas informações são armazenadas em meios digitais. O tamanho do universo digital cresce de forma exponencial. Segundo relatório publicado pela EMC Corporation¹ [22], em 2005, o volume de dados chegou a 130 exabytes; em 2010, superou 1 zettabyte e a previsão é que em 2015 chegue a quase 8 zettabytes [16].

Este universo digital citado em Gantz e Reinsel [16] expande e torna cada vez mais complexa a tarefa de filtragem de conteúdo relevante que atenda às preferências do usuário. Dentre as técnicas utilizadas na filtragem de conteúdo podemos citar os sistemas de recomendação[26].

¹<http://brazil.emc.com/index.htm?fromGlobalSelector>

Quando é preciso filtrar um grande conjunto de dados, podemos utilizar técnicas de recomendação para facilitar o processo de filtragem de informações relevantes e similares com as preferências do usuário. Para isto, é necessário informações sobre o indivíduo alvo da recomendação ou sobre o ambiente que influenciará na geração da recomendação. A partir daí, um sistema de recomendação poderá recomendar os itens que mais se aproximam das preferências ou características do usuário ou do seu ambiente [1] [19].

A partir da análise da literatura e dos sistemas de armazenamento em nuvem mais populares e utilizados atualmente, é notório que esses sistemas não fornecem ao usuário o serviço de recomendação de arquivos. Na maioria dos sistemas de armazenamento em nuvem, a filtragem de conteúdo é realizada por sistemas de busca, onde o usuário fornece termos-chaves e o sistema retorna arquivos com o título ou conteúdo similar aos termos apresentados pelo usuário.

Por outro lado, a enorme quantidade de sistemas de recomendação que funcionam como serviço em nuvem, não utilizam características da nuvem na geração de suas recomendações. Esses sistemas normalmente possuem como objetivo recomendar itens que atendam as preferências dos usuários, sem considerar requisitos do ambiente. Neste trabalho investiga-se a utilização de características da nuvem que podem ser usadas no processo de recomendação de arquivos em ambiente de armazenamento em nuvem.

2. FUNDAMENTAÇÃO TEÓRICA

O National Institute of Standards and Technology (NIST)² define computação em nuvem como um modelo que permite que um conjunto de recursos computacionais possam ser fornecidos sob demanda de forma a permitir que os mesmos sejam fornecidos e liberados rapidamente com o mínimo de esforço de gestão ou interação do fornecedor [11] [14] [23].

Entre os recursos disponíveis na tecnologia em nuvem, está o de armazenamento, o qual provê recursos e serviços de armazenamento baseados em servidores remotos que utilizam os princípios da computação em nuvem [32]. Armazenamento em nuvem tem duas características básicas: a primeira trata da infraestrutura da nuvem, a qual baseia-se em clusters de servidores baratos; a segunda tem o objetivo de, através dos *clusters* de servidores, armazenamento distribuído e redundância de dados, fazer múltiplas cópias dos dados armazenados para alcançar dois requisitos: alta escalabilidade e alta usabilidade [9]. A alta escalabilidade significa que o armazenamento em nuvem pode ser dimensionado para um grande aglomerado com centenas de nós ou peers de processamento. Alta usabilidade significa que o armazenamento em nuvem pode tolerar falhas de nós e que estas falhas não afetam todo o sistema [12].

2.1 Sistemas de Recomendação

De acordo com Lima et al. [2], Sistemas de Recomendação são ferramentas e técnicas de software que fornecem recomendação de itens [26]. As sugestões geradas por um SR podem estar relacionadas a processo de tomada de decisão em diversos contextos, como a escolha de itens em *e-commerces*, conteúdo similar as preferências do usuário, livros, amigos em redes sociais, rotas geográficas e etc [2] [28].

Uma recomendação pode se basear nas preferências de quem a faz e pode ser dirigida a um indivíduo específico, ou

para um público maior. Para a pessoa que recebe a recomendação, ela funciona como um filtro ou uma visão particular de um universo de possibilidades geralmente inacessível. Ela pode levar em consideração também a preferência de quem está à procura de sugestões e não apenas de quem a faz. É possível até mesmo fazer recomendação baseada nas opiniões de outras pessoas. Alguém que não é admirador do gênero Rock pode recomendar discos baseado no que seus amigos que apreciem tal estilo costumam ouvir. Ainda, a recomendação pode incluir explicações sobre como ela foi gerada para permitir que o seu recebedor a avalie [25] [26].

No universo dos sistemas de recomendação, existem diversas técnicas e formas de gerar recomendação, dentre quais destacam-se quatro técnicas como principais e mais utilizadas: recomendação baseada em conteúdo, filtragem colaborativa e filtragem híbrida [2] [6] [15] [26].

Filtragem baseada em conteúdo: esta categoria de sistemas recomenda ao usuário itens semelhantes àqueles em que ele demonstrou interesse no passado. Para tanto, o sistema analisa as descrições dos conteúdos dos itens avaliados pelo usuário para montar o seu perfil, o qual é utilizado para filtrar os demais itens da base. Esse conteúdo no qual ele se baseia são elementos explícitos como nome, descrição, *tags*, conteúdo, categorização ou *rating* do item a ser recomendado. Os resultados são o julgamento da relevância daqueles itens para o usuário, e a consequente recomendação ou não [24] [26]. Uma das vantagens de implantar o Filtro Baseado em Conteúdo é que a quantidade de usuários no sistema não interfere na eficácia do SR, já que se baseia somente no histórico do que o usuário já acessou. Em contrapartida, um sistema assim precisa de itens bem descritos, com informação suficiente para categorizá-los. Outro problema encontrado nesse tipo de recomendação é a sugestão de itens sempre muito parecidos, limitando os usuários de conhecer itens diferentes [26] [25].

Filtragem colaborativa: a abordagem de recomendação por filtragem colaborativa foi proposta inicialmente para suprir as deficiências da abordagem baseada em conteúdo. Com o passar dos anos, conquistou tamanha aceitação que hoje é provavelmente a técnica mais amplamente conhecida, implementada e utilizada para sistemas de recomendação [3] [17]. Na abordagem colaborativa, em contraste com a recomendação baseada em conteúdo, a compreensão ou conhecimento do conteúdo dos itens é totalmente prescindível. Ao invés de buscar itens disponíveis com conteúdos similares aos previamente avaliados positivamente pelo usuário para indicá-los, ela se apoia inteiramente na similaridade entre os usuários do sistema para o processo de sugestão. Partindo do princípio de que as melhores recomendações para um indivíduo são aquelas feitas por pessoas com preferências similares às dele, o sistema identifica estas pessoas para sugerir itens que as mesmas tenham aprovado e ainda não tenham sido consumidos pelo indivíduo [26] [27].

Filtragem híbrida: são sistemas de recomendação que utilizam duas ou mais técnicas, para amenizar os problemas apresentados por cada técnica [26] [27].

3. TRABALHOS RELACIONADOS

Existem alguns trabalhos na literatura que discutem e apresentam sistemas de recomendação em nuvem. Nesta seção, serão apresentados alguns SRs destacando o modelo de recomendação utilizado, objetivando avaliar as contribuições e diferenciais desta pesquisa.

²<http://www.nist.gov/srm/>

Lai et al. [20] que apresentam um sistema de recomendação de programas de televisão (TV) baseado em computação em nuvem e um *framework map-reduce*. Esta proposta de arquitetura tem como objetivo ofertar um *backend* escalável para suportar a demanda de processamento de dados em larga escala para um sistema de recomendação. No que tange os usuários, Lai et al. [20] os agrupam de acordo com suas preferências, cada programa de TV recebe um peso que é atribuído de acordo com o período de tempo que o usuário o assistiu. A popularidade de um programa é indicada pelo seu peso, os programas populares em um grupo de usuários, são recomendados para usuários de outros grupos que tenham semelhanças de preferências entre si. Nesta pesquisa os autores propõem a utilização de técnicas de computação em nuvem para lidar com grandes conjuntos de dados, devido ao seu poder computacional e de estrutura escalável.

Jung et al. [19], apresentam a plataforma CloudAdvisor de recomendação em nuvem. A proposta desta plataforma é recomendar configurações de nuvem de acordo com as preferências do usuário como orçamento, expectativa de desempenho e economia de energia para determinada carga de trabalho. Permitindo ainda, que o usuário faça comparação das recomendações recebidas, como qual é o melhor preço para a carga de serviço desejada. A plataforma tem como objetivo auxiliar o usuário na escolha dos melhores serviços e proporcionar aos provedores de serviço em nuvem à oportunidade de adequação às expectativas e preferências dos usuários.

Existem diversos sistemas de recomendação disponíveis na Internet, e boa parte destes sistemas estão relacionados à nuvem, seja como parte de sistemas em nuvem ou hospedados em servidores em nuvem. Muitos destes utilizam dados da nuvem em suas recomendações ou características da nuvem para gerar recomendações. A proposta desta pesquisa se diferencia das demais descritas nesta seção por utilizar características da nuvem na geração de suas recomendações. Desta forma, cada critério é parte da recomendação, que tem como objetivo proporcionar aos usuários a melhor utilização dos recursos em nuvem disponíveis.

4. O MODELO DE RECOMENDAÇÃO

O modelo de recomendação proposto neste trabalho, é composto por cinco critérios, que foram utilizados no processo de recomendação. Os critérios propostos foram definidos a partir da observação de sistemas de armazenamento em nuvem. Os critérios são: Similaridade, Disponibilidade, Taxa de *Download*, Tamanho do Arquivo e Popularidade do Arquivo.

Critério Similaridade: este critério atende ao requisito referente às preferências do usuário. Neste critério, é calculado a similaridade entre o conteúdo de um arquivo no qual o usuário tenha demonstrado preferência e arquivos armazenados em nuvem, que são candidatos a serem recomendados. Para calcular a similaridade entre os conteúdos dos arquivos, é proposto a utilização da técnica de similaridade do cosseno, que retorna um valor entre 0 (zero) e 1 (um). Esta abordagem, foi proposta por ser utilizada com frequência na avaliação de semelhança entre dois itens [4] [8] [21]. O cálculo de similaridade do cosseno é apresentado pela Equação 1:

$$St = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \quad (1)$$

Na Equação 1, é calculada a similaridade entre dois arquivos, o conteúdo de cada um é representando por um vetor (vetores A e B), de onde se obtém o produto de "A" e "B" e calcula-se a magnitude dos dois vetores. As magnitudes, são multiplicadas e divididas pelo produto escalar dos vetores "A" e "B". Os arquivos que sejam similares ao arquivo que represente as preferências do usuário serão ranqueados de acordo com o seu grau de similaridade, ou seja, quanto maior o *score* de similaridade do arquivo, melhor ranqueado ele será em referência aos demais similares as preferências do usuário. Por exemplo, caso o sistema de recomendação encontre dois arquivos "A" e "B", estes similares às preferências do usuário e com *score* de similaridade igual a "arquivo A = 0.8" e "arquivo B = 0.5". Neste cenário, o "arquivo A" será melhor ranqueado que o "arquivo B" no critério similaridade. A similaridade, torna-se imprescindível neste modelo de recomendação. Para tanto, este critério objetiva atender as preferências dos usuários em relação à filtragem de conteúdo relevante.

Critério Disponibilidade: refere-se ao tempo em que um arquivo estará disponível para o usuário. A disponibilidade, neste modelo, é medida em horas, ou seja a quantidade de horas em que um arquivo a ser recomendado está disponível na nuvem. Um arquivo só deve ser recomendado ao usuário se o mesmo estiver disponível e possibilitando o seu *download*. O critério disponibilidade representa uma das principais características e anseios quanto à tecnologia de computação em nuvem. A maioria dos usuários que migram para a nuvem são atraídos pela oferta de altas taxas de disponibilidade, elasticidade e mobilidade, que torna possível armazenar arquivos em grande quantidade e que estejam disponíveis e acessíveis a qualquer momento a partir da conexão com a Internet [7] [10] [23]. O cálculo do critério disponibilidade é apresentado na Equação 2:

$$D = h \cdot \left(\frac{1}{n}\right) \quad (2)$$

Na Equação 2, "D" é a quantidade horas em que um arquivo está disponível na nuvem, e "n" representa a quantidade de horas em que um arquivo pode ficar disponível na nuvem. Caso a nuvem fique disponível durante todo o dia, "n" será igual a 24 (vinte e quatro) horas. A quantidade de horas de disponibilidade é normalizada em um valor entre 0 (zero) e 1 (um). O exemplo a seguir, demonstra como o critério disponibilidade contribui para a geração de uma recomendação. Considere que dois arquivos "A" e "B" são similares, o arquivo "A" está disponível na nuvem no intervalo de tempo (14 às 16 horas), totalizando duas horas de disponibilidade. O arquivo "B" está disponível na nuvem de (14 às 18 horas), totalizando quatro horas de disponibilidade. Desta maneira, o arquivo que será melhor ranqueado é o arquivo "B", por estar disponível na nuvem por um tempo superior que o arquivo "A", permitindo o seu *download* em um espaço de tempo maior. O objetivo central é minimizar o risco do usuário não poder realizar o *download* e garantir que um arquivo recomendado esteja sempre acessível ao usuário.

Critério Taxa de *Download*: refere-se à taxa disponível para a realização do *download* de um arquivo na nuvem. O objetivo, é que arquivos que proporcionam melhores condições para a redução no tempo gasto no *download* sejam melhor ranqueados que os demais arquivos. A contribuição deste critério na redução do tempo gasto no *download* de um

arquivo recomendado, é produzida em conjunto com o critério “Tamanho do Arquivo”, apresentado no próximo item. Por exemplo, no caso de termos dois arquivos similares às preferências do usuário, onde o arquivo “A” tem o seu tamanho igual a 10 (dez) *Gigabytes*, e o arquivo “B” tem o seu tamanho igual a 2 (dois) *Gigabytes*. Neste cenário, o arquivo “A” será melhor ranqueado que o arquivo “B”, por proporcionar uma maior economia no tempo gasto em seu *download*. A taxa de *download* pode modificar o ranque de recomendações dependendo do momento em que a recomendação for calculada, principalmente em ambientes onde a taxa de *download* é oscilante. Este critério, tem valor de 0 (zero) a 3 (três) *Megabits* por segundo, este valor representa a media global de taxa de *downloads* apresentada pela Akamai³. Este critério é calculado pela Equação 3:

$$Td = ns \cdot \left(\frac{1}{n}\right) \quad (3)$$

Na Equação 3, a “Taxa de *Download*” é representada por “*Td*”, onde “*ns*” representa à taxa de *download* em Mbps, em seguida este valor é normalizado em um valor entre 0 (zero) e 1 (um), onde “*n*” representa o valor da media global de taxa de *downloads* em Mbps.

Critério Tamanho do Arquivo: este critério, corresponde ao tamanho do arquivo candidato a ser recomendado, tem como objetivo contribuir na tarefa de amenizar o tempo gasto no *download* de um arquivo recomendado. Como explicado, no item critério “Taxa de *Download*”, o critério “Tamanho do Arquivo”, está diretamente relacionado com o critério que mensura a taxa de *download* disponível. O ranque de recomendação, mudará de acordo com a taxa disponível para *download*. No caso, da taxa de *download* ser baixa, os arquivos com tamanho menores devem ser mais bem ranqueados que seus similares que são maiores. Da mesma forma, quando a taxa de *download* é alta, os arquivos com tamanhos maiores devem ser mais bem ranqueados. Exemplificando, o ranqueamento deste critério considerando que um arquivo “A” é similar ao arquivo “B”, o arquivo “A” têm o tamanho igual a 9 (nove) *gigabytes*. O arquivo “B”, tem tamanho igual a 2 (dois) “*gigabytes*”. Desta forma, o arquivo “B” será melhor ranqueado por apresentar melhores condições para a realização do seu *download* (menor tamanho), considerando que a taxa de *download* seja baixa. O cálculo deste critério é realizado pela Equação 4:

$$S = T \cdot \left(\frac{1}{n}\right) \quad (4)$$

Na Equação 4, o critério “Tamanho do Arquivo” é representado por “*S*”. O tamanho do arquivo, é medido em *gigabytes*, pelo fato de que boa parte dos sistemas de armazenamento em nuvem limita o tamanho máximo de um arquivo que pode ser salvo em nuvem e o espaço disponível para o usuário no sistema em *gigabytes*. O tamanho do arquivo é multiplicado por $\frac{1}{n}$, para que seja normalizado por um valor de 0 (zero) a 1 (um), o valor 1 (um) é dividido por *n* que é o tamanho máximo de um arquivo aceito no sistema de armazenamento em nuvem utilizado para implantação do modelo.

Critério Popularidade: este critério, representa a importância social de um arquivo na nuvem, avaliado por meio

³<http://www.akamai.com/>

Tabela 1: Pesos dos Critérios

Critério	Peso
Similaridade	4
Disponibilidade	2
Taxa de <i>Download</i>	2
Tamanho do Arquivo	1
Popularidade	1

da quantidade de *downloads* que foram realizados de um mesmo arquivo. Quanto maior é a quantidade de *downloads* realizados de um arquivo, maior será a popularidade desse arquivo na rede, resultando em um melhor ranqueamento do mesmo. A seguir um exemplo sobre o ranqueamento deste critério: um arquivo “A” é similar ao arquivo “B”, o arquivo “A” já teve dez *downloads* realizados, e o arquivo “B” já teve dezesseis *downloads* realizados. Desta forma, o arquivo “B” será melhor ranqueado, por obter um número maior de *downloads* efetuados no sistema de armazenamento em nuvem que o arquivo “A”. O cálculo, deste critério é representado pela Equação 5:

$$R = Qd \cdot \left(\frac{1}{n}\right) \quad (5)$$

Na Equação 5, o critério “Popularidade do Arquivo” é representado por *R*. A cada *download* realizado de um determinado arquivo, o contador de *downloads* desse arquivo é incrementado em 1 (um). Este valor, é medido de 0 (zero) a *n*, onde *n* é a maior quantidade de *downloads* realizados em um único arquivo no sistema. O valor de *n* é obtido na observação do histórico de *downloads* de arquivos no sistema de armazenamento em nuvem. No cálculo do critério, a quantidade de *downloads* de um arquivo *Qd* é normalizada, multiplicando *Qd* por $\frac{1}{n}$, desta forma o valor resultante deste critério será entre 0 (zero) e 1 (um).

4.1 Pesos dos Critérios

Em um mecanismo de recomendação, os critérios devem ser ponderados por pesos, para compor o *score* de recomendação, resultando em um *ranking* com os itens que devem ser recomendados ao usuário. A partir, da realização de testes de execução com diferentes pesos no cálculo de recomendação que será apresentado na próxima seção, propomos a utilização dos pesos descritos na Tabela 1. Os testes realizados tiveram como objetivo verificar a variação no resultado do *score* final de recomendação e quais pesos apresentaria a menor variação no *score*. Desta forma, os arquivos recomendados não apresentariam uma grande variação de similaridade com as preferências dos usuários.

Critério Similaridade: tem peso maior peso que os demais critérios, seu objetivo é garantir que o conteúdo de um arquivo recomendado ao usuário seja similar as suas preferências. Outro ponto motivador para este critério corresponder a 40% do *score* de recomendação é amenizar ou solucionar um dos principais problemas da técnica de recomendação baseada em conteúdo: a sugestão de itens sempre muito parecidos, limitando os usuários de conhecer novos conteúdos [26] [30]. Desta forma, o modelo de recomendação atenderá as preferências do usuário e ao mesmo tempo estará recomendando novos conteúdos que são relacionados às preferência do usuário. O critério Similaridade, é medido de 0 (zero) a 1 (um), um arquivo que possua similaridade

igual a 0 (zero) em comparação as preferências do usuário, somente será recomendado caso ele tenha uma alta taxa de popularidade na rede, e mesmo assim, o arquivo não será bem ranqueado em relação aos demais que apresentem alguma similaridade com as preferências do usuário.

Critério Disponibilidade: tem peso 2 (dois), por ser um dos critério mais importante no modelo proposto, representa o tempo em que um arquivo está disponível na nuvem, tornando possível o *download* de um arquivo recomendado. Este critério, é essencial para a recomendação de arquivos baseada em características da nuvem, por representar uma das principais característica e vantagens da utilização de sistemas de armazenamento em nuvem. O valor do critério Disponibilidade, será de 0 (zero) a 1 (um), um arquivo somente poderá ser recomendado ao usuário, se o mesmo estiver disponível.

Critério Taxa de Download: Este critério, terá o seu valor medido de 0 (zero) a 1 (um). Um arquivo que possua uma baixa taxa de *download* e seu tamanho seja maior que o tamanho dos demais arquivos similares a ele, o seu *score* de recomendação será menor, e consequentemente ele não será tão bem ranqueado quanto seus similares, por que o seu processo de *download* demandará mais tempo e processamento. Um arquivo com baixa taxa de download poderá aparecer no topo do *ranking* de recomendação, desde que seu tamanho seja proporcional à baixa taxa de *download*. Para que um arquivo na nuvem se torne recomendável, esse critério deve ser maior que 0 (zero), desta forma será possível realizar o *download* do arquivo.

Critério Tamanho do Arquivo: Este critério, tem peso inferior aos demais critérios, por não ser um critério crítico. Assim, um arquivo que tenha o tamanho igual ao máximo aceito pelo ambiente, poderá ser recomendado se a taxa de *download* for alta, garantindo bom desempenho no *download* do arquivo.

Critério Popularidade: Este critério, tem o seu peso inferior aos demais critérios, por não ser um critério crítico. Portanto, um arquivo que não seja popular na nuvem poderá ser recomendado ao usuário, o mesmo ocorre com os arquivos novos na rede, se o arquivo for bem ranqueado nos outros critérios do modelo.

5. AVALIAÇÃO

Para a avaliação do modelo proposto, foi implementado um protótipo como parte do sistema de armazenamento em nuvem Ustore⁴. Foram implementados, os critérios propostos no modelo RecCloud e o cálculo de recomendação. O Ustore, é uma ferramenta de armazenamento em nuvem baseada em uma arquitetura P2P híbrida que tem como objetivo armazenar dados com baixo custo e de forma que os mesmos não se tornem indisponíveis com eventuais problemas na rede [13] [29].

O mecanismo de recomendação do Ustore é baseado em conteúdo. No Ustore, as preferências dos usuários são representadas por arquivos que o usuário tenha adicionado em sua conta no sistema. O sistema de recomendação, apresenta recomendações baseadas na similaridade entre o arquivo de preferência do usuário e os arquivos armazenados na nuvem. Para que o usuário receba uma recomendação, ela deve ser solicitada a partir de um arquivo em sua conta.

5.1 Detalhes da Implementação

Critério Similaridade: a similaridade é calculada a partir do conteúdo dos arquivos, que são extraídos pelo Apache Lucene versão 3.6⁵ e o Apache Tika versão 1.2⁶.

Critério Disponibilidade: no Ustore, cada cliente possui um horário de funcionamento determinado inicialmente, que é utilizado para garantir a disponibilidade. Para chegar ao valor da taxa de disponibilidade de um cliente, subtraímos o tempo total possível para um cliente estar disponível em um dia (24 horas), pelo tempo em que o cliente ficou *on-line*. Desta forma, obtemos a quantidade de horas em que um cliente esteve disponível durante o dia

Critério Taxa de Download: é obtida a partir da observação do *download* de um arquivo qualquer, a partir de informações sobre o tempo gasto no *download* do arquivo e seu tamanho, chegamos à taxa de *download* da rede em “KBps” *Kilobits por segundo*. Para termos de avaliação, a taxa máxima de *download* utilizada foi a média global de taxa de *downloads* 3 (três) *Megabits* por segundo. Caso à taxa do usuário apresente um valor maior que a média global, o critério “Taxa de Download” será igual a 1 (um), representa que este critério recebeu o valor máximo no cálculo da recomendação.

Critério Tamanho do Arquivo: é obtido na base do Ustore em *KiloBytes* e convertido em *GigaBytes*. Para termos de avaliação da proposta, o tamanho máximo de um único arquivo no ambiente foi estabelecido em 10 *GigaBytes*. Desta forma, se um arquivo apresentar tamanho superior ao limite máximo estabelecido, este receberá o valor 1 (um), que representa a maior taxa ponderada do critério “Tamanho do Arquivo”.

Critério Popularidade do Arquivo: é representada pela quantidade de vezes em que foram realizados *download* de um arquivo. No Ustore, a quantidade de vezes em que foi realizado o *download* de um arquivo pode ser obtida diretamente na base de dados do sistema. Para termos de avaliação desta proposta, atribuímos a quantidade de 10 *downloads*, como a maior quantidade de *downloads* realizados de um único arquivo na rede.

5.2 Detalhes da Avaliação

Para avaliação deste trabalho, foi utilizada uma base de dados composta por 1.400 (mil e quatrocentos) artigos acadêmicos. Para a realização dos testes, foi necessário que os arquivos recebessem uma classificação que indicasse quais são os mais relevantes para o ranque de recomendação que será gerado. Para fazer essa classificação no conjunto de dados, foi utilizada a similaridade entre o contexto dos artigos com o conteúdo dos artigos utilizados para representar as preferências do usuário. Desta forma, todos os artigos que forem relacionados ao termo sistemas de recomendação, foram considerados relevantes para o ranque de recomendação. O resultado, foi de 462 (Quatrocentos e sessenta e dois) artigos foram considerados relevantes para esta avaliação, representando 33% da base de dados utilizada na realização dos testes.

Na avaliação, foi utilizada a metodologia proposta por Jain [18], onde é defendido que para realizar uma avaliação é preciso definir objetivos, métricas, fatores e níveis [22]. As métricas escolhidas para avaliar o desempenho do modelo

⁴<http://usto.re/>

⁵<https://lucene.apache.org/core/>

⁶<http://tika.apache.org/>

RecCloud, estão descritas a seguir:

Tempo gasto no Download: foram efetuados *downloads* dos arquivos recomendados e medido o tempo gasto para realizar cada *download*. Os resultados obtidos, serão comparados com o tempo gasto no *download* de arquivos recomendados utilizando o modelo de recomendação baseado em conteúdo do Ustore.

Precisão: é a taxa de itens relevantes recomendados no resultado. É dada, através da proporção entre o número de arquivos relevantes recomendados e o número total de arquivos recomendados [4] [21] [31]. O cálculo da métrica Precisão é apresentado na Equação 6

$$Precisao = \frac{|\{arqrelevantes\} \cap \{arqrecomendados\}|}{|\{arqrecomendados\}|} \quad (6)$$

Na Equação 6, “*arqrelevantes*” é a quantidade de arquivos recomendados que fazem parte do ranque de relevância, e “*arqrecomendados*” é a quantidade de arquivos recomendados para cada solicitação de recomendação. O resultado, é representado por valores entre 0 (zero) e 1 (um), quanto mais próximo de 1 (um), mais preciso é o sistema.

Recall: é a taxa de itens relevantes recomendados em relação à quantidade total de itens relevantes [21]. O cálculo da métrica *Recall* é apresentado na Equação 7.

$$Recall = \frac{|\{arqrelevantes\} \cap \{arqrecomendados\}|}{|\{arqrelevantes\}|} \quad (7)$$

Na Equação 7, “*arqrelevantes*” é a quantidade de arquivos recomendados que fazem parte do ranque de relevância, e “*arqrecomendados*” é a quantidade de arquivos recomendados para cada solicitação. O resultado é representado por valores entre 0 (zero) e 1 (um), quanto mais próximo de 1 (um) mais o sistema satisfaz a solicitação da recomendação.

F-measure: é a média ponderada das taxas de Precisão e *Recall*. O cálculo da métrica *F-measure* é apresentado na Equação 8

$$F - measure \alpha = \frac{(1 + \alpha) \cdot precisao \cdot recall}{(\alpha \cdot precisao) + recall} \quad (8)$$

Nesta avaliação, as taxas de Precisão e *Recall* têm o mesmo fator de importância. desta forma, o valor de α é igual a 1 (um). Logo, esta função só retornará um valor no intervalo entre 0 (zero) e 1 (um). Como parâmetros para comparação, foram consideradas satisfatórias taxas de Precisão de 0.40 e *Recall* de 0.42, resultados semelhantes obtidos por Blank et al. [5], Tanaka et al. [31] e Zhang et al. [33].

5.3 Cénarios de Avaliação

A avaliação deste trabalho, foi realizada em um sistema real de armazenamento em nuvem (Ustore). Na execução, foram montados dois cenários da seguinte forma:

Cenário I: O objetivo foi avaliar o desempenho do modelo RecCloud. Foram solicitadas recomendações para dez diferentes arquivos e avaliada a quantidade de arquivos recomendado para cada solicitação. Foram avaliadas as métricas Precisão, *Recall* e *F-Measure* observando os arquivos recebidos como recomendação. Os ranques de recomendação analisados, foram divididos em três níveis, o primeiro nível retornando 5 (cinco) arquivos, o segundo nível retornando 10 (dez) arquivos e nível três retornando 15 (quinze) arquivos como recomendação. Foram solicitadas recomendações

Tabela 2: Resultados do cenário I

	Nível I	Nível II	Nível III
Precisão	0.68	0.46	0.44
Recall	0.21	0.29	0.42
F-measure	0.32	0.35	0.42

para 10 (dez) artigos no modelo RecCloud, estas solicitações resultaram respectivamente em 50 (cinquenta), 100 (cem) e 150 (cento e cinquenta) artigos recomendados.

Cenário II: o objetivo foi medir o tempo gasto no *download* de arquivos recomendados e comparar os resultados com um modelo baseado em conteúdo. Com isso, foi possível avaliar se o modelo proposto atingiu um dos objetivos, que é amenizar o tempo gasto no *download* dos arquivos recomendados. Neste cenário, avaliamos a métrica “Tempo Gasto no *Download*”, foram solicitadas recomendações para 10 (dez) artigos diferentes no modelo RecCloud e no modelo baseado em conteúdo do Ustore, estas solicitações resultaram em 100 (cem) artigos recomendados, 50 (cinquenta) por cada modelo. Nesta métrica, foram utilizados ranques de recomendação com 5 (cinco) artigos retornados para cada solicitação de recomendação.

5.4 Resultados

Cenário I: Na Tabela 2, são apresentados os resultados obtidos nas taxas de Precisão, *Recall* e *F-measure*, estes resultados foram alcançados utilizando o modelo RecCloud para cada solicitação de recomendação e variando a quantidade de arquivos recomendados para cada solicitação nos níveis I, II e III.

A partir dos resultados apresentados na Tabela 2, foi observado que a melhor taxa de precisão obtida foi de 0.68 no “nível I”, onde foram retornados 5 (cinco) artigos para cada solicitação de recomendação. Os resultados obtidos na métrica de precisão foram superiores aos resultados utilizados como referência para termos de comparação. A melhor taxa de *recall* obtida foi de 0.42 no “nível III”, onde foram retornados 15 (quinze) artigos para cada solicitação de recomendação. A melhor taxa de *Recall* obtida nesta avaliação foi similar à taxa obtida no trabalho de Zhang et al. [33]. A partir das taxas de Precisão e *Recall*, foi calculada a taxa de *F-Measure*, onde foi obtido como melhor taxa o valor 0.42, este valor foi obtido no “nível III”.

Analisando os resultados apresentados, foi observado que o melhor resultado da taxa de Precisão foi obtido em nível diferente do nível onde foi obtida a melhor taxa de *Recall*, e que, a maioria dos arquivos relevantes recomendados estava no início dos ranques de recomendação. Este cenário, justifica-se pelo critério de similaridade, que representa 40% de cada recomendação. Desta forma, os artigos com maiores taxas de similaridade ficaram no início dos ranques e os artigos que mesmo tendo a sua taxa de similaridade baixa foram classificados como recomendáveis e foram recomendados no final dos ranques de recomendação.

A partir da comparação dos resultados apresentados no decorrer deste trabalho com resultados encontrados na literatura, como, por exemplo, os trabalhos de Blank et al. [5], Tanaka et al. [31] e Zhang et al. [33], pode-se concluir que a proposta apresentada por este trabalho obteve resultados satisfatórios na geração de recomendações em sistemas de armazenamento em nuvem, utilizando características do

Tabela 3: Resultados do cenário II

	Média	Máximo	Mínimo
RecCloud	900	2.200	400
RecUstore	1.150	2.700	150

ambiente associadas à técnica de recomendação baseada em conteúdo.

Cenário II: neste cenário, foram realizadas dez solicitações de recomendações para 10 (dez) artigos diferentes no modelo RecCloud e no modelo baseado em conteúdo do Ustore, estas solicitações resultaram em 100 (cem) artigos recomendados 50 (cinquenta) arquivos recomendados por cada modelo. Foi analisado, o tempo gasto no *download* de cada arquivo recomendado. Na Tabela 3, é apresentada uma comparação do resultado obtido em cada modelo (média, máximo e o mínimo de tempo gasto nos *downloads* realizados).

A partir dos resultados apresentados na Tabela 3 em milissegundos, foi observado que, o modelo proposto neste trabalho proporcionou redução no tempo gasto no *download* dos arquivos recomendados. A redução média de tempo gasto nos *downloads* foi de 207,06 milissegundos, o que representa uma redução de 17,8%.

6. CONCLUSÃO

A avaliação realizada neste trabalho obteve resultados próximos aos resultados encontrados na literatura sobre avaliação de sistemas de recomendação. Desta forma, é possível consideração como relevantes às contribuições apresentadas por esta pesquisa. O escopo deste trabalho, não abrange todas as possíveis características da nuvem e de sistemas de armazenamento em nuvem que podem ser utilizadas na geração de recomendações. Desta forma, como trabalhos futuros podem ser adicionados novos critérios ao modelo de recomendação, assim como, realizar a avaliação do modelo com outras técnicas de avaliação de sistemas de recomendação.

7. REFERÊNCIAS

- [1] L. M. d. L. A. V. F. P. R. T. e. A. C. S. Adriano de Oliveira Tito, Arley Ramalho Rodrigues Ristar. Recroute: Uma proposta de aplicativo para recomendação de rotas de nibus utilizando informas contextuais dos usuarios. *Anais do X Simp Brasileiro de Sistemas de Informas*, pages 218–223, 2014.
- [2] T. B. M. d. S. Alezy Oliveira Lima, Ricardo Alexandre Afonso. Plataforma pguide: um modelo de recomendação para usuarios ms. *Anais do X Simp Brasileiro de Sistemas de Informa*, pages 73 – 84, 2013.
- [3] A. Ansari, S. Essegai, and R. Kohli. Internet recommendation systems. *JOURNAL OF MARKETING RESEARCH*, 37(3):363–375, 2000.
- [4] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
- [5] I. Blank, L. Rokach, and G. Shani. Leveraging the citation graph to recommend keywords. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 359–362, New York, NY, USA, 2013. ACM.
- [6] R. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.
- [7] J. F. S. Carvalho. Um mapeamento sistematico de estudos em cloud computing. Master's thesis, Universidade Federal de Pernambuco (UFPE), 2012.
- [8] Y.-C. Chen, H.-C. Huang, and Y.-M. Huang. Community-based program recommendation for the next generation electronic program guide. *Consumer Electronics, IEEE Transactions on*, 55(2):707–712, 2009.
- [9] C. M. R. da Silva, J. L. C. da Silva, R. M. Melo, R. B. Rodrigues, L. R. Lucien, S. P. D. Melo, A. Colares, and V. C. Garcia. A privacy maturity model for cloud storage services. In *2014 IEEE 7th International Conference on Cloud Computing, Anchorage, AK, USA, June 27 - July 2, 2014*, pages 944–945, 2014.
- [10] C. M. R. da Silva, J. L. C. da Silva, R. B. Rodrigues, G. M. M. Campos, L. M. Nascimento, and V. C. Garcia. Security threats in cloud computing models: Domains and proposals. In *2013 IEEE Sixth International Conference on Cloud Computing, Santa Clara, CA, USA, June 28 - July 3, 2013*, pages 383–389, 2013.
- [11] C. M. R. da Silva, J. L. C. da Silva, R. B. Rodrigues, L. M. Nascimento, and V. C. Garcia. Systematic mapping study on security threats in cloud computing. *CoRR*, abs/1303.6782, 2013.
- [12] J. Deng, J. Hu, A. Liu, and J. Wu. Research and application of cloud storage. In *Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on*, pages 1–5, 2010.
- [13] R. F. A. F. J. G. V. T. F. Dur F. Assad. Ustore: A private cloud storage software system. In F. Daniel, P. Dolog, and Q. Li, editors, *Web Engineering*, volume 7977 of *Lecture Notes in Computer Science*, pages 452–466. Springer Berlin Heidelberg, 2013.
- [14] N. C. L. R. e Edmir Parada Vasques Prado. Caractericas dos servi de computa em nuvem usados por organizas brasileiras. *Anais do X Simp Brasileiro de Sistemas de Informa*, pages 661 – 673, 2013.
- [15] P. L. d. G. e. J. P. d. A. Fernando M. Figueira Filho. Um sistema de recomendação para fs de discussa web baseado na estimativa da expertise e na classifica colaborativa de conteudo. *Anais do V Simp Brasileiro de Sistemas de Informa*, pages 306 – 313, 2007.
- [16] J. Gantz and D. Reinsel. Extracting value from chaos state of the universe : An executivesummary., Junho 2011. 1-12.
- [17] J. L. Herlocker. *Understanding and improving automated collaborative filtering systems*. PhD thesis, University of Minnesota, 2000. AAI9983577.
- [18] R. K. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley, 1 edition, Apr. 1991.
- [19] G. Jung, T. Mukherjee, S. Kunde, H. Kim, N. Sharma, and F. Goetz. Cloudadvisor: A recommendation-as-a-service platform for cloud configuration and pricing. In *Services (SERVICES), 2013 IEEE Ninth World Congress on*, pages 456–463, 2013.
- [20] C.-F. Lai, J.-H. Chang, C.-C. Hu, Y.-M. Huang, and

- H.-C. Chao. Cprs: A cloud-based program recommendation system for digital tv platforms. *Future Gener. Comput. Syst.*, 27(6):823–835, June 2011.
- [21] S. Lee, D. Lee, and S. Lee. Personalized dtv program recommendation system under a cloud computing environment. *IEEE Trans. on Consum. Electron.*, 56(2):1034–1042, May 2010.
- [22] M. A. S. Machado. Uma abordagem para indexacao e buscas full-text baseadas em contedo em sistemas de armazenamento em nuvem. Master’s thesis, Universidade Federal de Pernambuco (UFPE), 2013.
- [23] P. Mell and T. Grance. The NIST Definition of Cloud Computing. Technical report, National Institute of Standards and Technology, Information Technology Laboratory, July 2009.
- [24] M. J. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27(3):313–331, 1997.
- [25] W. O. F. G. M. M. C. V. C. G. F. A. D. R. E. A. Ricardo Batista Rodrigues, Carlo M. R. da Silva. A cloud-based recommendation system. In B. White and P. Isaías, editors, *Proceedings of the IADIS International Conference WWW/Internet 2013*, pages 384 – 386, Fort Worth, Texas, USA, October 2013. 978-989-8533-16-6.
- [26] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.
- [27] R. B. Rodrigues. Reccloud: Um modelo de recomenda de arquivos para sistemas de armazenamento em nuvem. Master’s thesis, Universidade Federal de Pernambuco (UFPE), 2014.
- [28] R. B. Rodrigues, F. A. Dur V. C. Garcia, C. M. R. da Silva, R. R. Souza, and R. E. Assad. A cloud-based recommendation model. In *7th Euro American Conference on Telematics and Information Systems, EATIS’14, Valparaíso, Chile - April 02 - 04, 2014*, page 23, 2014.
- [29] M. S. P. J. T. G. V. A. R. Silva, A. Machado. Desenvolvendo aplicativos peer-to-peer (p2p) no contexto de data storage para ambientes de cloud computing. *Computa, S. B., editor, SBSI.*, 2013.
- [30] N. Stormer, H.; Werro and D. Risch. Recommending products with a fuzzy classification. *COLLECTeR Europe*, 2006.
- [31] M. G. M. Thiago Fujisaka Tanaka. Classifica de revises para constru de perfm sistemas de recomenda. In *Webmedia 2012 XVIII Simp Brasileiro de Sistemas Multima e Web*, 2012.
- [32] W. Zeng, Y. Zhao, K. Ou, and W. Song. Research on cloud storage architecture and key technologies. In *Proceedings of the 2Nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, ICIS ’09, pages 1044–1048, New York, NY, USA, 2009. ACM.
- [33] Z. Zhang, S. Shang, S. R. Kulkarni, and P. Hui. Improving augmented reality using recommender systems. In *Proceedings of the 7th ACM Conference on Recommender Systems, RecSys ’13*, pages 173–176, New York, NY, USA, 2013. ACM.

Sistema de Previsão do Tempo de Chegada dos Ônibus Baseado em Dados Históricos Utilizando Modelos de Regressão

Alternative Title: Prediction System of Bus Arrival Time Based on Historical Data Using Regression Models

Kássio R. Coquita
Centro de Informática -
Universidade Federal de
Pernambuco (UFPE)
50.733-970 – Recife/PE –
Brasil
krc3@cin.ufpe.br

Arley R. R. Ristar
Centro de Informática -
Universidade Federal de
Pernambuco (UFPE)
50.733-970 – Recife/PE –
Brasil
arr2@cin.ufpe.br

Adriano L. I. Oliveira
Centro de Informática -
Universidade Federal de
Pernambuco (UFPE)
50.733-970 – Recife/PE –
Brasil
alio@cin.ufpe.br

Patrícia C. A. R.
Tedesco
Centro de Informática -
Universidade Federal de
Pernambuco (UFPE)
50.733-970 – Recife/PE –
Brasil
pcart@cin.ufpe.br

RESUMO

Os Sistemas Inteligentes de Transporte são aplicações de tecnologias da informação e comunicação que visam à melhoria da área de transporte. O fornecimento de informações sobre a chegada do ônibus nas paradas é muito importante e útil aos passageiros e gestores de trânsito. Este artigo apresenta a proposta de um sistema de previsão do tempo de chegada do ônibus na parada onde o usuário está localizado. Para isso, um estudo experimental na linha de ônibus Campina do Barreto de número 722 em Recife-PE foi realizado, comparando o modelo de Máquina de Vetor de Suporte para Regressão (SVR) e a rede neural do tipo Máquina de Aprendizado Extremo (ELM) na estimativa do tempo gasto para o ônibus percorrer paradas adjacentes. Os experimentos foram realizados utilizando os dados de log de GPS de ônibus na Região Metropolitana do Recife, e os resultados mostraram que para este sistema o SVR obteve uma performance melhor, quando comparado com a ELM.

Palavras-Chave

Sistemas Inteligentes de Transporte, Previsão, Regressão, SVR, ELM.

ABSTRACT

Intelligent Transportation Systems are applications of information and communication technologies aimed at improving the transportation area. Providing information about bus arrival time on the bus stop is very important and useful to passengers and transit managers. This paper presents a proposed bus arrival time prediction system at the bus stop where user is located. For this, an experimental study on bus route named Campina do Barreto No. 722 in Recife-PE, was performed by comparing the regression model for Support Vector Machine (SVR) and the neural network Extreme Learning Machine (ELM) to estimate the time it takes to go through adjacent bus stops. The experiments

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.
Copyright SBC 2015.

were performed using the bus GPS log data in the metropolitan region of Recife, and the results showed that for this application the SVR significantly outperforms ELM.

Categories and Subject Descriptors

C.4 [Performance of Systems]: measurement techniques, modeling techniques, performance attributes; D.4.8 [Operating Systems]: Performance - modeling and prediction; G.3 [Probability and Statistics]: robust regression, statistical computing.

General Terms

Algorithms and Performance.

Keywords

Intelligent Transportation Systems, Prediction, Regression, SVR, ELM.

1. INTRODUÇÃO

O tráfego nas grandes cidades brasileiras tem piorado a cada dia que passa. Na realidade brasileira, os congestionamentos nas vias têm levado o transporte público a se tornar ineficiente, dado o tempo gasto para se locomover de um ponto a outro. O transporte público seria uma alternativa para a redução do fluxo de veículo nas vias se não fosse a baixa qualidade do serviço que as empresas oferecem. Um dos exemplos da baixa qualidade desse serviço é o longo tempo de espera que os passageiros têm que enfrentar nos pontos de ônibus.

De acordo com [3], esse tipo de problema contribui para a relutância dos usuários em utilizarem esse meio de transporte. Prover a informação do tempo de chegada do ônibus pode permitir que os passageiros planejem melhor suas viagens, diminua o tempo de espera e ansiedade nas paradas, além de poder diminuir o risco de assaltos na parada, já que o usuário só precisaria ir para o ponto de ônibus quando o mesmo estivesse se aproximando.

Em um cenário em que as informações sobre o trânsito são de grande utilidade para os usuários, a área de Sistema de Transporte Inteligente ganha importância e surge para ajudar a melhorar a vida das pessoas. Além disso, prover informações de grande relevância para os usuários de transporte coletivo urbano

(ônibus), melhorando a qualidade do serviço é mais uma contribuição desse tipo de sistema.

Neste artigo, foram utilizados os dados históricos coletados a partir de dispositivos GPS instalados nos ônibus da Região Metropolitana do Recife. A linha de ônibus Campina do Barreto de número 722 foi selecionada para treinamento e validação dos modelos de regressão escolhidos. Os dados históricos de GPS foram submetidos a um pré-tratamento para deixá-los com as informações desejadas (não contidas nos dados originais), e posteriormente, utilizá-los para criar os modelos de previsão com a Máquina de Vetor de Suporte para Regressão, ou *Support Vector Regression* (SVR) e com a Máquina de Aprendizado Extremo, ou em inglês, *Extreme Learning Machine* (ELM).

O restante deste artigo está organizado da seguinte forma: na Seção 2 são discutidos alguns trabalhos relacionados, já na Seção 3 e 4 são introduzidos os conceitos que cerceiam este projeto. Na Seção 5, é descrita a proposta do sistema. A descrição sobre os dados e o seu processo de pré-tratamento é realizada na Seção 6. A configuração do experimento é exposta na Seção 7. Os resultados preliminares obtidos com a implementação dos modelos de regressão podem ser vistos na Seção 8. Por fim, na Seção 9 é apresentada as conclusões e perspectivas para o trabalho futuro.

2. TRABALHOS RELACIONADOS

Foram encontrados na literatura alguns trabalhos que calculavam a previsão de tempo de chegada dos ônibus nas paradas, tais como em [5], que na tentativa de realizar previsões sobre o tráfego de maneira mais eficiente, estendeu vários métodos de previsão de séries temporais existentes capazes de lidar com os dados gravados em intervalos de tempo irregulares. Essa característica permitiu que os autores trabalhassem com dados de fonte intermitentes, tais como o Sistema de Posicionamento Global (GPS). Os resultados mostram que o melhor desempenho em termos da média absoluta do erro relativo foi obtido por meio de um modelo de rede neural.

Em [3] foi desenvolvido um algoritmo de previsão de aprendizagem não-supervisionada baseado em um modelo de dados históricos. Essa base histórica continha dados de localização e velocidade dos ônibus obtidos através do sensor de GPS instalados neles. Esses dados foram utilizados para treinar uma Rede Neural do tipo *backpropagation* para prever a velocidade média e o tempo de chegada em determinadas seções da estrada. Os resultados experimentais mostraram que o algoritmo proposto obteve uma melhor precisão na previsão, quando comparado com outros métodos baseados em dados históricos do tempo de viagem.

No estudo de [6], foi apresentada uma abordagem que combinava dados históricos e em tempo real para prever o tempo de chegada dos ônibus. Esse trabalho foi dividido em duas fases, em que na primeira, uma rede neural com Função de Ativação de Base Radial foi utilizada para aprender e aproximar a relação não-linear em dados históricos. Já na segunda fase, o Filtro de Kalman foi aplicado para ajustar a previsão realizada na primeira etapa utilizando dados *online*. Os resultados obtidos pela combinação desses modelos foram comparados com a Regressão Linear Múltipla, Rede Neural do tipo *backpropagation* e Rede Neural com Função de Ativação de Base Radial sem o ajuste *online*. Essa comparação mostrou que a abordagem proposta no estudo obteve um desempenho melhor na predição quando comparada aos outros modelos.

Assim, pôde-se observar na literatura consultada, um evidente e crescente interesse no que concerne ao tema proposto neste projeto. Esse levantamento bibliográfico também permitiu um maior entendimento das técnicas, ajudando na elaboração de estratégias para utilização das mesmas, ou até mesmo, na etapa de pré-processamento dos dados, de modo a contribuir efetivamente para o desenvolvimento deste trabalho.

Além disso, diante dos trabalhos citados, não foi encontrada uma pesquisa que aplicasse as Máquinas de Vetor de Suporte para Regressão ou as Máquinas de Aprendizado Extremo na previsão do tempo gasto pelo ônibus para percorrer paradas adjacentes, bem como, alguma que utilizasse a mesma característica dos dados para que pudesse ser realizada uma comparação mais detalhada entre elas e este trabalho em relação à performance de predição.

A presente pesquisa foi conduzida seguindo os conceitos referentes aos Sistemas Inteligentes de Transporte e aos métodos de regressão: Máquina de Vetor de Suporte para Regressão e Máquina de Aprendizado Extremo, conforme será discutido nas próximas seções.

3. SISTEMAS INTELIGENTES DE TRANSPORTE

Os Sistemas Inteligentes de Transporte, ou em inglês, *Intelligent Transportation Systems* (ITS) são aplicações de tecnologia para a área de transporte que integram informação, métodos de comunicação e tecnologias, a fim de auxiliar o sistema de transporte de uma região, integrando pessoas, estradas e veículos [1].

Um dos principais objetivos dos ITS é monitorar o tráfego das vias para otimizar as viagens, fornecendo suporte à mobilidade dos passageiros e veículos, de modo a aumentar a eficiência e a segurança nas redes de transportes atuais.

Uma das aplicações dos ITS é controlar os horários de chegada dos ônibus nas paradas, tendo em vista que o tempo de chegada é influenciado por vários fatores, como por exemplo, condições meteorológicas, o congestionamento do tráfego, as catástrofes naturais, e etc. Tais fatores resultam em atrasos no cronograma e inconveniência para os passageiros devido ao tempo de espera nas paradas de ônibus [4].

Neste artigo, é proposta uma abordagem na área de Sistemas de Transporte Inteligente para previsão do tempo de chegada do ônibus nas paradas, fazendo uso de técnicas de regressão e tendo como base um conjunto de dados históricos de logs de GPS.

4. OS MÉTODOS DE REGRESSÃO

Em problemas de regressão é dado um conjunto de treinamento $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathcal{X} \times \mathbb{R}$, onde $\mathcal{X}$ denota o espaço dos padrões de entrada, por exemplo, $\mathbb{R}^d$. O principal objetivo da regressão é encontrar uma função $f(x)$ que melhor modele esses dados de treinamento [7]. No caso deste trabalho, há o interesse em construir e comparar modelos de regressão para prever o tempo gasto para o ônibus percorrer paradas adjacentes tendo como base os dados históricos.

4.1 Máquina de Vetor de Suporte para Regressão

Um dos modelos utilizados nesta pesquisa foi o ε -SVR, tipo de Máquina de Vetor de Suporte para Regressão, que define a função de perda ε -insensível. Esse tipo de função define uma

margem, ou faixa, em torno dos valores reais, como pode ser visto na Figura 1.

A principal ideia desse método é que os erros menores que certos limiares $\varepsilon > 0$ sejam ignorados. Ou seja, erros gerados por pontos que estejam na faixa do tubo são considerados zero. Em contrapartida, erros causados por pontos localizados fora da faixa são medidos pelas variáveis ξ e ξ^* , como mostrado na Figura 1.

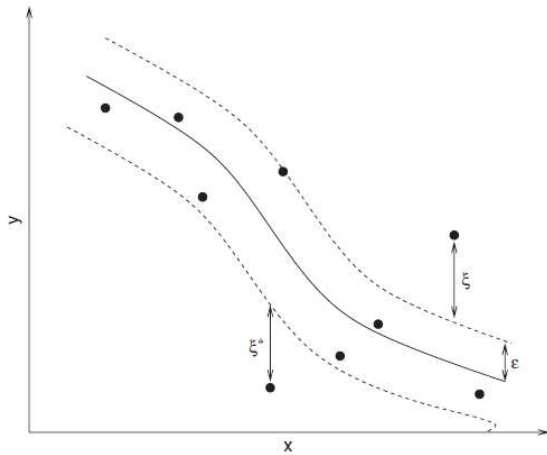


Figura 1. Regressão usando ε -SVR

4.2 Máquina de Aprendizado Extremo

Outro método para regressão adotado nesta pesquisa foi uma rede neural conhecida como Máquina de Aprendizado Extremo, ou em inglês, *Extreme Learning Machine* (ELM). A principal característica desta rede está na sua capacidade de aprendizado sem a necessidade de iterativa configuração dos parâmetros, ou seja, suas características são ajustadas aleatoriamente.

A ELM mostra que as camadas ocultas são importantes, mas podem ser ajustadas aleatoriamente e de forma independente do conjunto de dados de treinamento. Ao contrário dos métodos convencionais de aprendizagem que devem ter contato com os dados de treinamento antes de gerar os parâmetros dos neurônios ocultos, a ELM pode gerar aleatoriamente os parâmetros dos neurônios ocultos antes de ver os dados de treinamento [2].

Quando comparados os esforços desses métodos em relação à configuração, a ELM possui uma maior simplicidade, uma vez que requer como parâmetros apenas os números de neurônios e a função de ativação. Já o SVR necessita, em geral, dos parâmetros de custo (C), o nível de precisão da função aproximada (ε) e a largura (γ), em caso de kernel RBF.

5. A PROPOSTA DO SISTEMA

O sistema proposto terá a função de informar ao usuário o tempo de chegada do ônibus, tendo como entrada a linha do veículo, a parada em que o passageiro espera o veículo, o tipo do trajeto (ida ou volta), se a requisição está sendo realizada na hora de pico (07:00 às 09:00 e 17:00 às 19:00), e se é um dia de semana, além da informação da localização do ônibus desejado.

A Figura 2 mostra um diagrama esquemático do trajeto de um ônibus. A rota é dividida em segmentos s entre paradas adjacentes e i é o número da parada ($i = 1, 2, 3, \dots, n$).

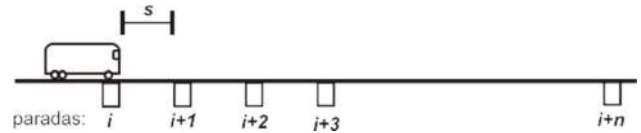


Figura 2. Esquema de uma rota de ônibus com várias paradas.

Quando um usuário requisita o tempo de chegada de um ônibus até a sua posição, o sistema irá verificar a localização do veículo desejado e contar o número de paradas entre o ônibus e o usuário. Caso o ônibus não esteja estacionado em uma parada, o sistema preverá o tempo que falta para ele atingir a próxima parada $P_{inicial}$ e somará com os tempos previstos que o veículo levará da parada $P_{inicial}$ para a parada P_{final} . O esquema funcional do sistema pode ser visto pela Figura 3.

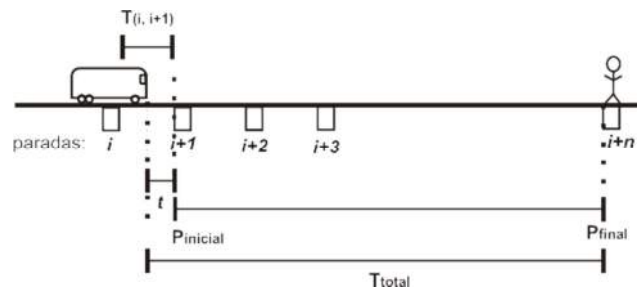


Figura 3. Esquema do sistema de previsão de tempo de chegada do ônibus.

Portanto, o tempo de chegada do ônibus na parada em que o usuário está localizado é então calculado de acordo com a Equação (1):

$$T_{total} = \sum_{i=P_{inicial}}^{P_{final}} (T_{(i,i+1)}) + t \quad (1)$$

Na qual:

T_{total} = é o tempo total previsto que o ônibus levará da sua localização até chegar na parada P_{final} ;

$T_{(i,i+1)}$ = é o tempo previsto que o ônibus leva da parada i até a próxima parada ($i + 1$);

t = é o tempo que falta para o ônibus atingir a parada mais próxima.

Assim, fazendo uso do modelo de regressão apropriado para prever o tempo que o ônibus levará para percorrer paradas adjacentes, o sistema poderá usá-lo para a calcular o tempo total que o veículo consumirá para chegar até o usuário.

6. DADOS

O estudo utilizou os dados de GPS de ônibus da linha Campina do Barreto de número 722, fornecidos pela empresa Grande Recife Consórcios e Transporte (GRCT). Os dados foram coletados durante 8 dias, no período de 16 de agosto de 2012 até 23 de agosto de 2012. A linha Campina do Barreto (722), que é realizada pela empresa São Paulo, possui um trajeto que vai do bairro Cajueiro até o ponto de retorno no Cais de Santa Rita, em Recife. A escolha dessa linha de ônibus foi motivada por ser uma linha mais central na cidade do Recife, ter uma frequência de saída um pouco menor que a média (menos pontos coletados) e a linha possuir alguns trechos com ida e volta juntas (passam pela mesma rua em sentido contrário) e outros diferentes.

A empresa GRCT disponibilizou um arquivo de texto com os dados de todas as linhas e de todos os ônibus da região metropolitana do Recife, que são enviados para um servidor pelo equipamento de GPS instalado nos ônibus. Esse arquivo, chamado nesta pesquisa de arquivo-fonte, são *dumps* de um banco de dados e cada informação é armazenada entre ponto e vírgulas numa ordem específica. Uma linha desse arquivo-fonte pode ser visto a seguir:

- SPA;722;541;2012-08-21;00:00:04.0000000;-8,022242;-34,88397;0;2012-08-21 00:00:11.7370000

A ordem das informações no formato acima pode ser constatada de acordo com as seguintes posições: (I) sigla da empresa de ônibus; (II) o número da linha; (III) o prefixo do veículo; (IV) a data da informação armazenada no banco de dados; (V) o horário que a informação foi armazenada no banco de dados; (VI) a latitude; (VII) a longitude; (VIII) a velocidade instantânea e (IX) o *timestamp* de aferição do dispositivo.

Vale destacar que o formato das coordenadas (latitude e longitude) utilizado por este trabalho é o padrão WGS 84.

Para o presente trabalho, será utilizado o *timestamp* de aferição do dispositivo, presente na posição IX, como o atributo de tempo. Os campos IV e V serão ignoradas, pois devido aos problemas de conexão para o envio das informação ao servidor de banco de dados, a data e hora que a informação foi armazenada difere da data e hora de aferição.

Nesse arquivo-fonte estão contidos quase 49 milhões de registros referentes a todas as linhas de ônibus e ao período citado acima. No entanto, nem todas as informações contidas no arquivo foram necessárias para o uso neste trabalho, uma vez que apenas os registros referentes à linha 722 foram utilizados.

Além desse arquivo-fonte, foram usados dados das localizações das paradas de ônibus. Essas localizações também foram fornecidas pela empresa de transporte e serão importantes para dividir o trajeto da linha de ônibus escolhida em segmentos. Para a linha de ônibus escolhida, foram contabilizadas 55 paradas, das quais 31 fazem parte do trajeto de ida e 24 de volta.

Assim, diante dessa quantidade de dados disponibilizada foi fundamental realizar um pré-processamento neles, para selecionar as informações necessárias, formatá-las, e remover as possíveis inconsistências contidas no arquivo.

6.1 Pré-processamento dos Dados

Nesta seção serão descritos os procedimentos para seleção, formatação e inserção dos registros no banco de dados para poderem ser utilizados. Além disso, serão explicados os procedimentos de obtenção das viagens realizadas pelos ônibus da linha 722 selecionada para este estudo.

Esta etapa é fundamental para o trabalho, pois nem todas as informações contidas no arquivo-fonte são necessárias, e também, pela possibilidade de encontrar inconsistências, uma vez que o projeto faz uso de dados reais. Além disso, esta seção de pré-processamento serve para explicar em qual banco de dados as informações tratadas serão armazenadas e qual aplicação utilizar para poder visualizá-las.

Conforme mencionado na seção anterior, as informações do arquivo-fonte que serão necessárias neste trabalho são: (I) a sigla da empresa de ônibus; (II) o número da linha; (III) o prefixo do veículo; (VI) a latitude; (VII) a longitude, e (VIII) a velocidade

instantânea. Vale lembrar que as informações contidas na posição IV e V não serão utilizadas.

Com a definição das informações necessárias, a primeira etapa do pré-processamento consistiu em selecioná-las no arquivo-fonte e armazená-las em um arquivo-intermediário.

Esse resultado intermediário foi armazenado em um banco de dados espacial, para poder melhor manipular os dados resultantes. O SGBD escolhido foi o PostgreSQL, por ser uma ferramenta de *software* livre e possuir uma boa extensão para trabalhar com dados espaciais (PostGIS).

De acordo com as informações selecionadas, foi necessário criar tabelas para armazenar (a) as empresas de ônibus, (b) os ônibus que cada empresa possui, (c) as linhas de ônibus, e (d) o histórico de registro. Essa tabela de histórico armazena as informações dos dados que estão contidas no arquivo-intermediário, com colunas para a linha do ônibus, o ônibus que realizou o trajeto, a velocidade instantânea do ônibus, o *timestamp* e a geolocalização em formato POINT do PostGIS. Não foi criada uma chave primária ou identificador único para esta tabela, porém existem chaves estrangeiras para as linhas e os ônibus. As duas chaves estrangeiras em combinação com o *timestamp* são teoricamente únicos, mas, como esses dados são reais, é possível que haja ruídos entre eles.

De fato, alguns ruídos foram encontrados nos dados armazenados, pois como o envio de informações para o servidor não obedece a um intervalo regular de tempo, devido aos problemas de conexão, foi necessário remover os registros duplicados.

Após essa primeira fase, os dados da tabela histórico referentes à linha de número 722 do dia 16 de agosto de 2012 até 23 de agosto de 2012, foram selecionados e copiados para uma tabela temporária. A criação dessa tabela temporária se dá na tentativa de obter uma melhor performance no momento da utilização dos dados, já que do total de registros, apenas 58.694 foram obtidos com essa seleção.

É válido ressaltar que uma linha de ônibus pode ser executada por vários ônibus, e a linha 722 conta com uma frota de 6 ônibus para realizar o trajeto.

Os dados desses ônibus da frota da linha 722 podem ser visualizados na Figura 4, em que é possível verificar o trajeto formado pelos pontos espaciais referentes à linha 722. A ferramenta utilizada para gerar essa figura foi a Quantum GIS (ou QGIS), ferramenta com licença de software livre e que se integra muito bem ao PostgreSQL.

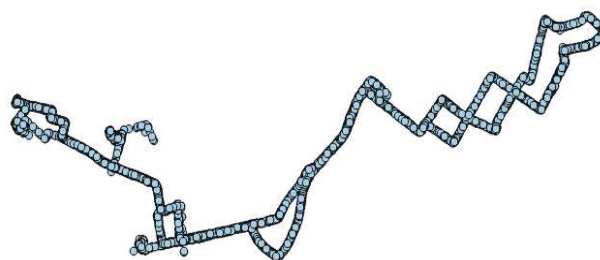


Figura 4. Visualização dos pontos de GPS do trajeto da linha 722.

Já na Figura 5, os dados podem ser visualizados utilizando a ferramenta Google Maps, com destaque para o trajeto de ida (em vermelho) e o de volta (em cinza).

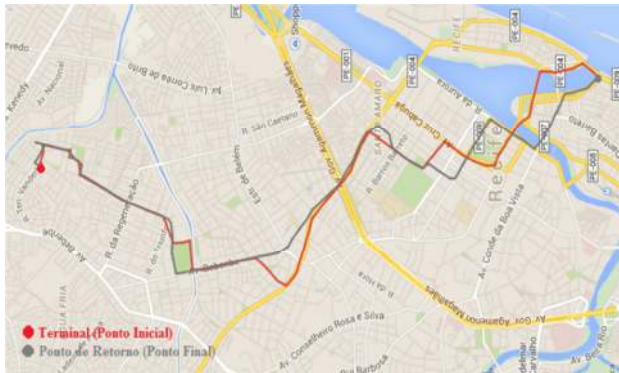


Figura 5. Visualização da linha 722 pelo Google Maps.

Posterior ao processo de seleção dos dados históricos dos logs dos ônibus, foi necessário submetê-los a alguns procedimentos, de modo a obter as informações desejadas (fatores de interesse) que se adequem às necessidades do sistema proposto descrito na Seção 5.

Portanto, com o objetivo de extrair desses métodos as previsões do tempo que o ônibus levará para percorrer paradas adjacentes, foram considerados os seguintes fatores de interesse: parada i , parada $i + 1$, tipo de trajeto, hora de pico, se é dia de semana, e o tempo de viagem entre as paradas, sendo esse último a variável dependente. Um exemplo desses fatores de interesse pode ser visto na Tabela 1.

Tabela 1. Exemplo dos fatores de interesses da linha 722.

Parada (i)	Parada (i + 1)	Tipo de Trajeto	Hora de Pico	Dia de Semana	Tempo de Viagem (seg)
1	2	ida	sim	sim	25.2
2	3	ida	sim	sim	48.0
3	4	ida	não	não	120.0
20	21	volta	sim	sim	78.0
21	22	volta	não	não	234.0
22	23	volta	não	não	30.0

Para poder chegar à esses fatores de interesse foi de suma importância identificar as coordenadas dos pontos iniciais e finais do trajeto de ida e volta, ou seja, onde estão localizados o terminal e o ponto de retorno da linha de ônibus escolhida.

Essa informação está disponível na página de consulta de itinerários no site da empresa fornecedora dos dados utilizados. Dessa forma, esses dados foram adquiridos manualmente e aplicados como pontos de referência para a divisão dos trajetos entre as viagens.

Após a determinação do ponto inicial e final da rota, foram identificadas todas as paradas de ônibus que estavam no trajeto de ida ou volta desta linha, conforme pode ser visto na Figura 6. Essas paradas de ônibus foram usadas para separar em partes menores os trajetos, e com isso obter os segmentos listados na Tabela 1. Com a realização desse processo, foram extraídos um total de 10.662 segmentos para a ida e volta.

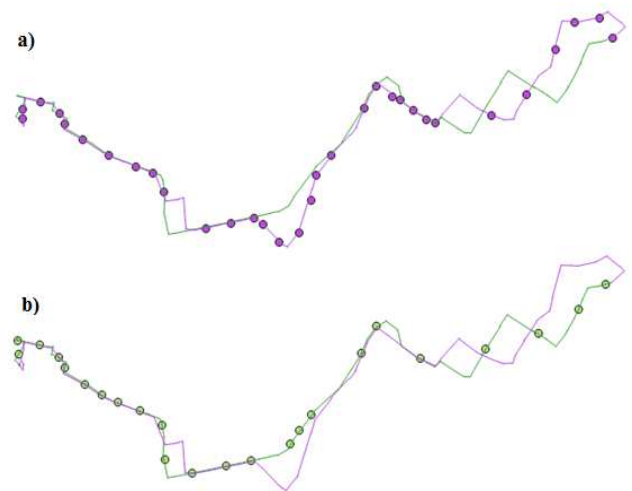


Figura 6. a) paradas de ônibus do trajeto de ida; b) paradas de ônibus do trajeto de volta.

É preciso mencionar que o tempo gasto por um ônibus para percorrer as paradas adjacentes foi calculado a partir do *timestamp* de pontos de GPS do ônibus. Como esses pontos de GPS nem sempre têm exatamente a mesma localidade das paradas, foi necessário ajustar o tempo que o ônibus gastaria para percorrer o segmento. A Figura 7 exibe o esquema realizado para calcular o tempo de viagem de um segmento (entre as paradas i e $i + 1$).

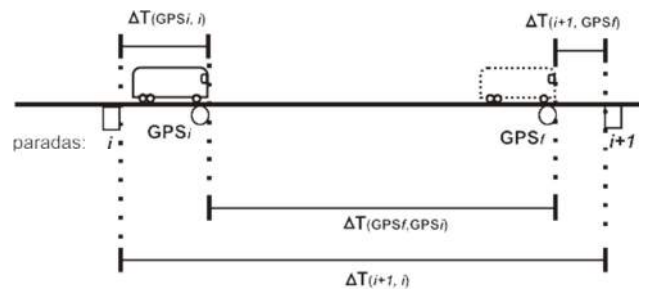


Figura 7. Esquema do cálculo do tempo de viagem de um segmento.

Portanto, o tempo que um ônibus leva para percorrer um segmento, definido entre as paradas i e $i + 1$ é então calculado de acordo com a Equação (2):

$$\Delta T_{(i+1,i)} = \Delta T_{(GPS_f, GPS_i)} + \Delta T_{(GPS_i, i)} + \Delta T_{(i+1, GPS_f)} \quad (2)$$

Na qual:

GPS_i = é o ponto de GPS inicial que o ônibus está localizado;

GPS_f = é o ponto de GPS final que o ônibus está localizado;

$\Delta T_{(i+1,i)}$ = é o tempo de viagem entre duas paradas adjacentes realizadas pelo mesmo ônibus;

$\Delta T_{(GPS_f, GPS_i)}$ = é a diferença de tempo que um ônibus leva para percorrer os pontos de GPS final e inicial;

$\Delta T_{(GPS_i, i)}$ = é a diferença de tempo que um ônibus leva para percorrer a parada i e seu ponto de GPS inicial;

$\Delta T_{(i+1, GPS_f)}$ = é a diferença de tempo que um ônibus leva para percorrer o seu ponto de GPS final e a parada $i + 1$.

Assim, diferentemente das informações contidas no arquivo-intermediário, foram obtidos os fatores de interesse independentes, tais como: as paradas i e $i + 1$, o tipo do trajeto, se a medição foi realizada no horário de pico, e se a medição foi obtida em dia de semana, além do fator dependente, que é o tempo de viagem entre duas paradas adjacentes (em segundos).

Vale lembrar que o horário de pico considerado na pesquisa é das 07:00 às 09:00, pela manhã, e das 17:00 às 19:00 à noite. Além disso, esses fatores e seus respectivos valores, conforme mostrado na Tabela 1, serão chamados de dados históricos do tempo de viagem para cada segmento entre as paradas de ônibus adjacentes.

7. CONFIGURAÇÕES DO EXPERIMENTO

A regressão realizada neste trabalho utilizou o conjunto de dados históricos do tempo de viagem, mencionado na seção anterior. Este conjunto de dados possui cinco variáveis independentes – parada i , parada $i + 1$, tipo de trajeto, hora de pico e se é dia de semana – e uma variável dependente – tempo de viagem entre duas paradas adjacentes.

Neste projeto, o objetivo principal é estimar o tempo de viagem do ônibus, em segundos, baseado em dados históricos.

O tempo de viagem do ônibus em um determinado segmento (entre as paradas i e $i + 1$) pode ser denotado como $T_{(i, i+1)}$, sendo $1 \leq i \leq n$, para o ônibus que está percorrendo a rota através da parada i ($i = 0, 1, 2, 3, \dots, n$). Dessa forma, $T_{(i, i+1)}$ pode ser obtido através das variáveis: parada i , parada $i + 1$, tipo de trajeto, hora de pico e se é dia de semana.

As simulações dos algoritmos SVM do tipo “eps-regression” e ELM para regressão foram realizadas utilizando as bibliotecas LibSVM e elmNN do R, respectivamente. Além disso, os dados utilizados foram normalizados entre 0 e 1, conforme indicado pelas bibliotecas.

Vale destacar que a medida de performance utilizada neste trabalho é a média percentual do erro absoluto, ou em inglês, *mean absolute percentage error* (MAPE). A MAPE é dada pela Equação (3), onde $\{x_i\}$ é a observação real, $\{\hat{x}_i\}$ é o valor previsto e N é o valor do número de dados.

$$MAPE = \frac{100}{N} \times \sum_{i=1}^N \left| \frac{x_i - \hat{x}_i}{x_i} \right| \quad (3)$$

Para estimar os parâmetros ideais para o modelo, foi utilizada a validação cruzada do tipo k -fold em combinação com o *grid search*. Na validação cruzada k -fold, o conjunto de dados é particionado em k subconjuntos mutuamente exclusivos, onde k é o número total de subconjuntos. O conjunto de treinamento é construído utilizando $k - 1$ subconjuntos para estimar os parâmetros, e o subconjunto restante é utilizado para validação ou teste do modelo. Este processo é repetido k vezes, onde a cada iteração são empregados diferentes dados para previsão. Neste estudo foi utilizado um k igual a 10.

O método de *grid search*, foi aplicado na variação dos valores dos parâmetros necessários ao SVR e ELM. Nas simulações realizadas para o SVR foram considerados os parâmetros C , ε e γ para o kernel RBF. Nesse método, os parâmetros assumiram os seguintes valores: $C = \{10, 100, 1000\}$, $\varepsilon = \{10^{-3}, 10^{-4}, 10^{-5}\}$ e $\gamma = \{1, 10^{-1}, 10^{-2}\}$. Dessa forma, considerando as diferentes combinações para os parâmetros, foram executadas 27 simulações com o kernel RBF.

Já nas simulações utilizando o método ELM foram variados os parâmetros referentes às funções de ativação e aos números de neurônios. Nesse método os parâmetros assumiram os seguintes valores: função de ativação = *{symmetric hard-limit, base radial}*, nº de neurônios = $\{10, 15, 20, 25, 30, 35, 40, 45, 50\}$, além de um k -fold igual a 10. Logo, considerando as diferentes combinações para os parâmetros, foram executadas 18 simulações no total. Vale salientar que para cada combinação de parâmetros da ELM a validação cruzada (10 k -folds) foi executada 30 vezes.

8. ANÁLISE COMPARATIVA DOS RESULTADOS

A análise dos resultados produzidos pelos métodos de regressão foi realizada pela comparação entre as medidas de performance (MAPE) obtidas para cada modelo.

A Tabela 2 compara as medidas de performance para o SVR considerando o kernel RBF, e variando os parâmetros C , ε e γ . Já a Tabela 3, compara o desempenho dos parâmetros para a rede ELM, alterando os valores do número de neurônio e da função de ativação.

Ainda na Tabela 2 é possível verificar que os valores da MAPE ao longo da alternância dos parâmetros não variou muito. Para os resultados obtidos com o ELM, exibidos na Tabela 3, é possível verificar que a MAPE para função de ativação *symmetric hard-limit* aumentou à medida que o número de neurônios também foi aumentando. Além disso, o mesmo aconteceu com a MAPE para o ELM com função de ativação de base radial.

Tabela 2. Comparação do desempenho do SVR de acordo com a variação dos parâmetros

Método/Parâmetros	MAPE	Desvio Padrão
(C = 10, $\varepsilon = 10^{-3}$, $\gamma = 1.0$)	3.514	0.857
(C = 10, $\varepsilon = 10^{-3}$, $\gamma = 10^{-1}$)	2.484	1.331
(C = 10, $\varepsilon = 10^{-3}$, $\gamma = 10^{-2}$)	2.101	0.646
(C = 10, $\varepsilon = 10^{-4}$, $\gamma = 1.0$)	3.246	1.049
(C = 10, $\varepsilon = 10^{-4}$, $\gamma = 10^{-1}$)	2.470	1.373
(C = 10, $\varepsilon = 10^{-4}$, $\gamma = 10^{-2}$)	1.958	0.434
(C = 10, $\varepsilon = 10^{-5}$, $\gamma = 1.0$)	3.103	1.007
(C = 10, $\varepsilon = 10^{-5}$, $\gamma = 10^{-1}$)	2.390	1.211
(C = 10, $\varepsilon = 10^{-5}$, $\gamma = 10^{-2}$)	2.045	0.499
(C = 100, $\varepsilon = 10^{-3}$, $\gamma = 1.0$)	2.624	1.713
(C = 100, $\varepsilon = 10^{-3}$, $\gamma = 10^{-1}$)	2.710	1.822
(C = 100, $\varepsilon = 10^{-3}$, $\gamma = 10^{-2}$)	2.159	0.629
(C = 100, $\varepsilon = 10^{-4}$, $\gamma = 1.0$)	2.564	1.762
(C = 100, $\varepsilon = 10^{-4}$, $\gamma = 10^{-1}$)	2.677	1.879
(C = 100, $\varepsilon = 10^{-4}$, $\gamma = 10^{-2}$)	2.344	0.772
(C = 100, $\varepsilon = 10^{-5}$, $\gamma = 1.0$)	2.576	1.598
(C = 100, $\varepsilon = 10^{-5}$, $\gamma = 10^{-1}$)	2.771	2.204
(C = 100, $\varepsilon = 10^{-5}$, $\gamma = 10^{-2}$)	2.341	0.885
(C = 1000, $\varepsilon = 10^{-3}$, $\gamma = 1.0$)	5.716	11.657
(C = 1000, $\varepsilon = 10^{-3}$, $\gamma = 10^{-1}$)	2.698	1.720
(C = 1000, $\varepsilon = 10^{-3}$, $\gamma = 10^{-2}$)	2.564	1.547
(C = 1000, $\varepsilon = 10^{-4}$, $\gamma = 1.0$)	5.729	11.559
(C = 1000, $\varepsilon = 10^{-4}$, $\gamma = 10^{-1}$)	2.827	1.794
(C = 1000, $\varepsilon = 10^{-4}$, $\gamma = 10^{-2}$)	2.552	1.526
(C = 1000, $\varepsilon = 10^{-5}$, $\gamma = 1.0$)	5.797	11.633
(C = 1000, $\varepsilon = 10^{-5}$, $\gamma = 10^{-1}$)	2.812	1.893
(C = 1000, $\varepsilon = 10^{-5}$, $\gamma = 10^{-2}$)	2.533	1.489

Tabela 3. Comparação do desempenho do ELM de acordo com a variação dos parâmetros

Método/Parâmetros	MAPE	Desvio Padrão
(nh = 10, fun = symmetric hard-limit)	6.39	2.16
(nh = 15, fun = symmetric hard-limit)	7.38	2.32
(nh = 20, fun = symmetric hard-limit)	8.50	2.31
(nh = 25, fun = symmetric hard-limit)	9.75	2.56
(nh = 30, fun = symmetric hard-limit)	9.86	2.23
(nh = 35, fun = symmetric hard-limit)	10.48	1.90
(nh = 40, fun = symmetric hard-limit)	11.35	1.75
(nh = 45, fun = symmetric hard-limit)	12.52	3.27
(nh = 50, fun = symmetric hard-limit)	12.88	2.04
(nh = 10, fun = radial basis)	8.35	1.85
(nh = 15, fun = radial basis)	10.97	1.23
(nh = 20, fun = radial basis)	13.25	3.27
(nh = 25, fun = radial basis)	14.56	3.30
(nh = 30, fun = radial basis)	19.02	4.67
(nh = 35, fun = radial basis)	25.31	9.98
(nh = 40, fun = radial basis)	43.17	19.93
(nh = 45, fun = radial basis)	85.74	38.99
(nh = 50, fun = radial basis)	153.63	122.06

Os resultados mostraram que o melhor valor da MAPE para o SVR foi de 1,958%, em que a combinação de parâmetros foi: $C = 10$, $\varepsilon = 10^{-4}$ e $\gamma = 10^{-2}$. Já para o ELM, o melhor resultado da MAPE foi de 6,39%, quando combinada a função de ativação *symmetric hard-limit* e o número de neurônios igual a 10.

Como o SVR obteve o melhor desempenho na previsão, um segundo experimento foi realizado, refinando seus parâmetros com base nos melhores valores alcançados anteriormente. Esses parâmetros assumiram os seguintes valores: $C = \{10, 50, 90\}$, $\varepsilon = \{10^{-3}, 5 \times 10^{-4}, 5 \times 10^{-5}\}$ e $\gamma = \{10^{-2}, 5 \times 10^{-2}, 10^{-3}\}$.

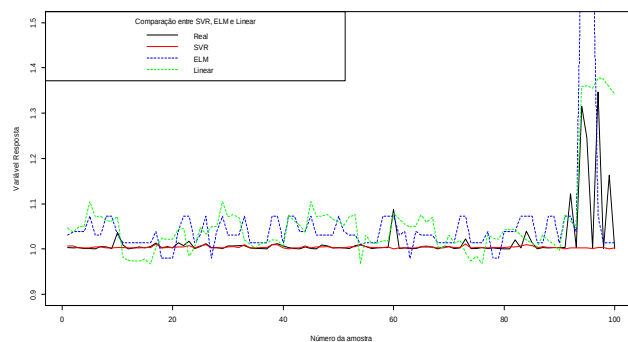
Esse segundo experimento foi executado na tentativa de alcançar melhores resultados dos que já se tinham para o SVR. Ele foi realizado sob os mesmos procedimentos do experimento anterior, ou seja, com uma validação cruzada do tipo *k-fold*, com $k = 10$ e um *grid search* variando os novos valores dos parâmetros C , ε e γ .

Com a realização desse novo experimento, a MAPE do SVR foi reduzida de 1,958% para 1,492%, em que a combinação de parâmetros foi: $C = 50$, $\varepsilon = 10^{-3}$ e $\gamma = 10^{-3}$.

Após esse processo de determinação dos parâmetros, o SVR e o ELM foram reconstruídos utilizando os parâmetros que geraram o melhor desempenho para cada método. Além disso, um modelo clássico de regressão linear foi implementado pela função “lm” do R, para comparar os resultados. Essa comparação com o modelo linear se faz necessária, para saber se um método de regressão mais simples conseguiria obter uma performance melhor na previsão.

Para o treinamento e teste, foram selecionados 75% e 25% dos dados, respectivamente. Com essa execução dos métodos, a medida de desempenho (MAPE) obtida para cada método foi de 1,009% para o SVR, 4,499% para o ELM e 4,844% para o Linear.

Na Figura 8, parte da amostra foi selecionada para exibir um contraste entre os dados reais e as previsões realizadas por cada método.

**Figura 8. Comparação entre os métodos SVR, ELM e Linear para previsão**

Assim, comparando os resultados das medidas de desempenho entre os métodos SVR, ELM e Linear aplicados para prever o tempo que o ônibus levará para percorrer paradas adjacentes, é possível verificar que o SVR obteve uma performance superior ao ELM e ao modelo clássico de regressão Linear.

9. CONCLUSÃO

O presente trabalho avaliou a performance de dois modelos para previsão do tempo de chegada do ônibus na parada. Foram investigados os métodos de Máquina de Vetor de Suporte para Regressão (SVR), bem como *Extreme Learning Machine* (ELM). Tais métodos utilizaram, para prever o tempo de chegada, as variáveis independentes: identificação das paradas adjacentes (i e $i + 1$), o tipo do trajeto, se a medição foi realizada no horário de pico e se é dia de semana. Essas informações foram extraídas à partir de dados históricos de GPS de ônibus cedidos pela empresa de transporte público do Recife, Grande Recife Consórcios e Transporte.

Os resultados de cada método foram comparados levando em consideração a média percentual do erro absoluto, ou MAPE. De acordo com as simulações, o método SVR obteve uma média percentual de 1.009%, que é menor que as médias de 4.499% e 4.844% obtidas pelos modelos ELM e Linear, respectivamente.

Para versões futuras deste trabalho, ficará a possibilidade de expandir e melhorar o sistema proposto. Uma das primeiras possibilidades a ser adicionada ao projeto, poderia ser a implementação da funcionalidade para calcular o tempo de espera do ônibus nas paradas. Outra possibilidade futura está na utilização da velocidade instantânea para melhorar a precisão da previsão. Além disso, características inerentes a linha do ônibus como, por exemplo, a quantidade de sinais de trânsito e quantidade de cruzamentos na via, podem ser consideradas.

10. AGRADECIMENTOS

Este trabalho foi apoiado pelo INES - Instituto Nacional de Ciência e Tecnologia para Engenharia de Software (<http://www.ines.org.br/>), financiado pelo CNPq, processo 573964/2008-4.

11. REFERÊNCIAS

- [1] An, S.-H., Lee, B.-H., Shin, D.-R. 2011. A Survey of Intelligent Transportation Systems. Communication Systems and Networks (CICSyN), 2011 Third International Conference, pp.332-337.

- [2] Huang, G., Wang, D. H., Lan, Y. 2011. Extreme learning machines: a survey, em: *International Journal of Machine Learning and Cybernetics*, vol. 2, issue 1, p. 107-122.
- [3] Pan, J., Dai, X., Xu, X., Li, Y. 2012. A Self-Learning Algorithm for Predicting Bus Arrival Time Based on Historical Data Model. *Cloud Computing and Intelligent Systems (CCIS)*, IEEE 2nd International Conference, vol. 3, Hangzhou, p. 1112-1116.
- [4] Raut, R. D., Goyal, V. K. 2012. Public transport Bus Arrival Time Prediction with Seasonal and Special Emphasis on Weather Compensation changes using RNN, em: *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 1, issue 6.
- [5] Ye, Q., Szeto, W. Y. e Wong, S. C. 2012. Short-Term Traffic Speed Forecasting Based on Data Recorded at Irregular Intervals, em: *Intelligent Transportation Systems*, vol. 13, no. 4.
- [6] Wang, L., Zuo, Z. e Fu J. Bus Arrival Time Prediction Using RBF Neural Networks Adjusted by Online Data, 9th International Conference on Traffic & Transportation Studies, (2014), p. 67-75.
- [7] Oliveira, A. L. I. Estimation of Software Project Effort with Support Vector Regression, *Neurocomputing* 69, (2006), p. 1749–1753.

Avaliação do comportamento de consumidores no processo de decisão de compra no M-Commerce e no E-Commerce

Alternative Title: Evaluation of Consumer Shopping Behavior in M-Commerce and E-Commerce

Felipe Lemos

Pontifícia Universidade Católica de Minas Gerais
Contagem - Minas Gerais
CEP: 32010-025
felipe.lemos.almeida@gmail.com

Luís Fabrício Góes

Pontifícia Universidade Católica de Minas Gerais
Contagem - Minas Gerais
CEP: 32010-025
lfgoes@yahoo.com.br

RESUMO

Este artigo avalia o comportamento de consumidores que efetuam compras através do Comércio Eletrônico utilizando uma plataforma Web e outra Mobile. Pretende-se analisar o motivo pelo qual consumidores preferem concluí-las através de um computador, já que o mercado de dispositivos móveis se encontra em alta.

Palavras-Chave

Comércio Eletrônico, Compartamento do Consumidor, Mobile Commerce, Compras, Internet, Loja Virtual

ABSTRACT

This paper intends to analyze why *e-commerce* is more popular than mobile commerce among internet users, despite the growth of mobile market. Through survey research carried out among 130 users, it was possible to infer that people use mobile devices along the day and prefer using computers than mobile devices in all phases of their purchases. The possible reasons of that behavior is related to safety conditions of m-commerce and errors that occur on electronic transactions. Mobile devices are used auxiliary tool to check available products on market and to follow the track of requests already done.

Categories and Subject Descriptors

K.4.4 [Electronic Commerce]: Miscellaneous; J.4 [Social and Behavioral Sciences]: Economics

General Terms

Verification

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

Keywords

m-commerce, e-commerce, consumer behavior, online store, online shop

1. INTRODUÇÃO

O Brasil, atualmente, é um dos países mais promissores quando o assunto gira em torno do Comércio Eletrônico. Isso ocorre devido às inovações no setor de Tecnologia da Informação e Comunicação e o aumento significativo de pessoas com acesso a Internet no país. Pesquisas revelam que no ano de 2013 foram faturados 28 bilhões com Comércio Eletrônico, cerca de 25% a mais que no ano anterior [8]. Diante de um cenário tão promissor, empresas estão utilizando a Internet para oferecer variados tipos de serviços. Dentre eles, o *E-Commerce* vem se destacando por apresentar, nos últimos anos, alto índice de crescimento.

Uma nova tendência que também aparece no mercado brasileiro, chamando atenção das empresas, é o *M-Commerce*, Comércio Eletrônico para dispositivos móveis. O surgimento desse tipo de serviço se deu em razão do crescimento nas vendas de *smartphones*, que, por sua vez, foi um dos fatores que impulsionaram o índice de pessoas com acesso à Internet no Brasil [23].

Considerando o exposto anteriormente, as operadoras de telefonia foram levadas a evoluir o serviço de Internet e melhorar sua infraestrutura para suportar a referida demanda. Isso ocasionou durante os últimos anos a evolução da Internet de 2G, para 3G, e, atualmente para 4G. Assim, com o aumento da velocidade da Internet e a praticidade em acessá-la, vêm colaborando para o desenvolvimento do *M-Commerce*, juntamente com número de consumidores para essa plataforma. As corporações, visando explorar esse mercado, começam a migrar suas lojas para essa linha. Igualmente, dados mostram a evolução nas vendas através do *M-Commerce*, chegando a 4% do total arrecadado com Comércio Eletrônico no ano de 2013 [8].

No entanto, esse percentual está abaixo do esperado, levando em consideração a quantidade de aparelhos conectados à Internet, que no mesmo ano foram registrados 96,1 milhões de acessos através de celulares, incluindo os *smartphones* [23].

Diante da situação, é possível afirmar que as condições são favoráveis para expansão do *M-Commerce* no Brasil, porém,

aproximadamente 86% das pessoas utilizam os *smartphones* apenas para buscar informações, preferindo, contudo, finalizar a compra através do *E-Commerce* ou na própria loja física [12].

Sendo assim, o principal objetivo deste trabalho é analisar o comportamento de consumidores no *M-Commerce*, fazendo a comparação com os que utilizam *E-Commerce* e desta forma será possível entender porque estes não finalizam suas compras pelo próprio dispositivo móvel.

A estrutura do artigo é dada pela seguinte organização: na seção 2 é feita uma abordagem sobre o *E-Commerce* e seus conceitos, e na sequência, é apresentado o *M-Commerce* na seção 3 juntamente com sua atuação no país 3.1. Na seção 4 é apresentado o comportamento do consumidor e as etapas do processo decisório de compra. A seção 5 contempla os trabalhos relacionados e na seção 6 é apresentada a metodologia. Os resultados da pesquisa são exibidos na seção 7 e para finalizar a conclusão e os trabalhos futuros aparecem na seção 8.

2. E-COMMERCE

Na atualidade, existem diferentes conceitos sobre o Comércio Eletrônico, entretanto, no presente trabalho, iremos abordar sua definição como qualquer tipo de transação comercial efetuada por empresas de forma eletrônica, a fim de fornecer algum tipo de serviço/produto [1]. Segundo Carvalho [5], “essa forma de negociar – o Comércio Eletrônico – engloba um conjunto de atividades que vai além da compra e venda de produtos e/ou serviços, mas que inclui pesquisa de mercado, oferta e promoção de produtos, suporte a clientes, dentre outras”.

Existem três modelos principais de e-business (negócios efetuados por meio eletrônico), de acordo com conceitos elaborados, pelos autores Deitel e Obrien[7, 18]; então, podemos citar: B2B (Business-to-Business), C2C (Consumer-to-Consumer) e B2C (Business-to-Consumer). No primeiro caso, B2B, é a negociação eletrônica envolvendo empresas, ao passo que o C2C consiste na transação on-line realizada entre pessoas físicas; em relação ao B2C, empresa para consumidor, será o modelo contemplado neste trabalho devido a um dos seus objetivos, que é o estudo do comportamento de usuários nas compras realizadas através de uma loja virtual.

3. M-COMMERCE

Conforme as considerações sobre *E-Commerce*, antes de se fazer uma abordagem sobre o *M-commerce*, é necessário dispor sobre o surgimento deste serviço, para que seja entendido porquê o mesmo, nos dias atuais, já é considerado a tecnologia que poderá substituir o *E-Commerce*.

O *M-Commerce*, Comércio Eletrônico para dispositivos móveis, teve início quando os celulares, por volta do ano 1990, começaram a ter acesso à Internet através da rede 1G, tecnologia que já existia desde os anos 80. Entretanto, muitos problemas eram enfrentados pelos usuários, que quase nunca conseguiam acessar a rede, devido a vários fatores, tais como: dificuldade de acesso, baixa velocidade (9.6kbps), e aparelhos que não ofereciam esse tipo conexão. Para suprir as limitações da Internet de primeira geração, as operadoras foram levadas a evoluir sua infraestrutura, sendo que na mesma década de 90, surgiu a segunda geração, a Internet 2G, essa por sua vez já era digital, possuía mais segurança

e uma velocidade de 20kbps a 50kbps [16].

Antes de dar continuidade na evolução da Internet móvel, é importante salientar que, em paralelo os celulares também estavam em constante evolução e, que tudo isso somente estava acontecendo porque a procura estava sendo muito maior do que a oferta, contudo a capacidade dos aparelhos era muito restrita e serviam somente para receber/fazer ligações, enviar mensagens e realizar algumas operações básicas. Quando esses foram adquirindo novos hardwares, consequentemente, adquirindo novas funções e utilidades, como jogos, possibilidade de escutar música, tirar fotos, entre outras, os usuários sentiram a necessidade de acesso à Internet, para comprar ou fazer download destes itens. Com evolução da Internet, juntamente com aumento de acessos proporcionados pelos celulares, empresas começaram a disponibilizar serviços/produtos para serem consumidos através deste novo meio.

Com aumento considerável de usuários, a Internet Móvel 2G evoluiu para 2,5G e posteriormente para 2,75G, essa já possuía velocidade de acesso aproximadamente de 400kbps, onde já era possível fazer download de jogos, músicas e comunicar com pessoas, de forma mais eficiente. As companhias Google e PayPal foram fundadas, por volta de 1998, e durante os 4 anos subsequentes Hotmail, Netscape, Yahoo e Amazon, ou seja, no ano de 2000, a Internet já funcionava como um comércio, onde consumidores buscavam, eletronicamente, produtos e faziam transações [14].

Outro fator que impulsionou o aumento de usuários na Internet através de celulares, juntamente com evolução dos mesmos para se tornarem *smartphones*, foram as redes sociais. Atualmente a Internet se encontra na sua quarta geração (4G), entretanto a maioria dos usuários ainda está na terceira geração (3G), tendo em vista que cerca de 80% utiliza esta conexão.

É possível concluir que, assim como o *E-Commerce*, o *M-commerce* possui fatores semelhantes que fizeram com que esse serviço pudesse ganhar força no mercado de Comércio Eletrônico, dentre eles, podemos citar: evolução da Internet móvel e evolução dos celulares.

3.1 M-Commerce no Brasil

O *M-commerce* vem aumentando seu percentual de vendas no Brasil, entretanto, ainda avança a passos curtos. Para Migliorini[16], alguns dos fatores a serem considerados consistem no fato das empresas não investirem neste serviço, pois, ainda trata-se de uma tecnologia inovadora. O autor cita também que a falta de interesse por parte dos varejistas se deve à segurança do consumidor. A falta de padrões em celulares, navegadores de Internet e a falta de investimento na estrutura de rede por parte das operadoras, também contribuem para que as organizações não invistam no *M-commerce*.

Já o consumidor, centro das atenções neste trabalho, também enfrenta muitas dificuldades ao usufruir deste serviço, dentre elas as restrições de hardware (tamanho da tela), acesso a rede Internet (muito lenta na maioria das vezes), adaptação (usuários não executam as mesmas tarefas que conseguiriam usando o computador), dificuldade (habilidade na execução da compra), todos esses são fatores que influenciam na hora da compra, principalmente através do celular [16]. Mas, apesar destes fatores, os consumidores, possuem uma boa perspectiva em relação a este mercado, uma vez que a comodidade é um fator de peso, que traz benefícios

para os mesmos, já que eles conseguem efetuar uma compra, ou realizar um pedido, em qualquer lugar e a qualquer hora sem precisar ligar o computador.

As vendas no Brasil chegaram a 4% do total arrecado com Comércio Eletrônico em 2013, sendo que no ano anterior foi de aproximadamente 2%. Um progresso considerado bom por especialistas da área, que ainda afirmam que até 2018, o *M-Commerce* já vai ter superado o *E-Commerce*.

4. COMPORTAMENTO DO CONSUMIDOR

Entender o comportamento do consumidor é de extrema importância para saber como ele fundamenta suas decisões na hora de efetuar uma compra. Segundo o modelo apresentado pelos autores Kotler e Keller[13], este processo é constituído por cinco etapas: reconhecimento das necessidades, procura de informação, avaliação das alternativas, decisão de compra e avaliação pós-compra. Como citado anteriormente, este estudo faz-se bastante necessário principalmente quando falamos de vendas online, pois o vendedor não tem um contato direto com o comprador. Com isso, é necessário que empresas que pratiquem vendas online entendam como serão as atitudes tomadas pelos e-consumidores, para que desta forma, as mesmas possam construir e/ou aprimorar suas estratégias de marketing.

O Kotler e Keller[13] modelo apontado anteriormente é utilizado por vários autores quando há necessidade de se realizar um estudo sobre o comportamento do consumidor na hora da compra, pois as etapas contidas nele demonstram todo o processo decisão, desde o início quando surge a necessidade até o final, que é o pós-compra. Neste trabalho foi utilizado este mesmo modelo nas plataformas a serem estudadas – *E-Commerce*, *M-commerce* – apresentando também conceitos de alguns autores sobre as etapas envolvidas no processo decisório.

4.1 Reconhecimento das necessidades

Esta primeira etapa é definida por Blackwell[9] como o início do processo de tomada de decisão, e esta pode ser causada por fatores como: mudança de circunstâncias, aquisição de produtos, consumo de produto, diferenças individuais e marketing.

Para Kotler[13], a necessidade para efetuar uma compra, pode surgir através de estímulos internos ou externos. No primeiro caso, são as necessidades normais (fome, sede, sexo), no momento que se tornam conscientes acabam se tornando um impulso, e, para o segundo caso, os desejos são causados por fatores externos, ou seja, algo que irá chamar atenção do consumidor e influenciá-lo a comprar.

Levando em consideração os conceitos apresentados anteriormente, empresas que trabalham com Comércio Eletrônico necessitam de informações sobre o que pode estimular uma compra online, pois o processo de decisão se inicia neste estágio, ou seja, vem da necessidade em adquirir/comprar algo, perceber como isso ocorre é um fator fundamental para o progresso no modelo B2C.

4.2 Procura de Informação

Após o surgimento da necessidade, ocorre o segundo estágio que é a busca por informações. Esta etapa consiste em realizar uma análise das informações que estão armazenadas na memória, bem como encontrá-las através de uma pesquisa [4].

Kotler e Keller[13] definem a busca de informações a partir

de níveis de interesse. Eles classificam estes em dois níveis; quando a busca é moderada é definido como atenção elevada, por outro lado, quando a procura é intensa, classifica-se como busca ativa de informações. Citam também, que as fontes de informações estão divididas em quatro grupos:

- Fontes pessoais: Família, amigos, conhecidos.
- Fontes Comerciais: Propaganda, vendedores, mostruários.
- Fontes públicas: Internet.
- Fontes Experimentais: Uso do produto [13].

Para Costa[6], o mercado disponibiliza todas as informações específicas dos produtos para que os consumidores possam consultá-las antes de efetuar uma compra, este processo é chamado de pesquisa pré-compra, ainda segundo o autor, o outro processo é denominado de pesquisa contínua, esse por sua vez ocorre quando o consumidor faz pesquisas constantes para se manter atualizado.

Na etapa anterior, o consumidor percebe a necessidade em adquirir algo, mas é nesta que ele irá buscar informações sobre o que deseja comprar, para confirmar se o produto será capaz de suprir tal necessidade. Esta tarefa é de extrema importância quando a compra será realizada através do Comércio Eletrônico, pois em ambos, a descrição, as características e as informações técnicas do produto, devem estar bem descritas. De acordo com pesquisa realizada pela HiMidia[11], a Internet é uma grande influenciadora na hora de decidir a compra. Ainda acrescenta em sua pesquisa que: 77% dos entrevistados procuram informações de produtos e serviços na Internet.

É possível destacar vantagens e desvantagens entre a busca de informações realizada através do computador e a realizada pelo celular. Um estudo, “Our Mobile Planet” encomendado pela Google e realizado pelas empresas MediaCT e TNS Intratest, revela que 46% dos usuários que utilizaram o celular para realizar uma pesquisa, acabaram não efetuando a compra e 27% finalizaram diretamente no *smartphone*, acrescenta também que, 45% dos usuários finalizaram a compra em um computador e outros 30% preferiram finalizar a compra em uma loja física[20].

Referente à pesquisa, é possível destacar que a possibilidade de acessar o *M-commerce* a qualquer hora e em qualquer lugar, dá a esta uma vantagem sobre seu concorrente (*E-Commerce*). Entretanto, o mesmo pode apresentar problemas quanto à navegabilidade, agilidade e usabilidade.

4.3 Avaliação das alternativas

Após obter informações necessárias sobre o que deseja, o consumidor, a partir de uma análise prévia, irá escolher onde irá de fato efetuar sua compra. Para Blackwell[4], existem alguns critérios que são levados em consideração na hora da compra, como: preço, marca, país de origem, etc. Já Silva[22] complementa dizendo que é nesta fase em que se avaliam os riscos da compra, e somente se o usuário não encontrar nada que o atrapalhe, o mesmo dará prosseguimento no processo.

Blackwell[4] afirma que os consumidores devem levantar um conjunto de considerações a serem feitas, quais fatores devem ser determinantes para compra, ou então escolher através do seu próprio conhecimento. Angelo e Silvera[2]

também ressaltam que os mesmos devem efetuar comparações entre as alternativas capazes de solucionar tal necessidade. É importante ressaltar que neste estágio, o consumidor irá definir onde comprar e o que comprar, de acordo com as necessidades consideradas de importância [6].

Segundo Kotler[13], o consumidor avalia um produto por sua quantidade de atributos, e cada um sua capacidade de proporcionar benefícios, aos quais possa suprir suas necessidades. Complementa dizendo que os consumidores prestarão mais atenção aos atributos que são capazes de atender o que está sendo procurado.

Ao chegar nesta etapa, o mesmo pode decidir comprar através de uma loja virtual ou até mesmo em uma loja física, isso vai depender dos fatores que este considerar relevantes para tomar sua decisão. Angelo e Silvera[3] ressaltam alguns fatores que levará o consumidor a escolher a loja virtual, como por exemplo: conveniência, economia de tempo, preços mais baratos, variedade de produtos oferecida, acesso 24 horas, mais informações sobre os produtos, personalização, acesso a produtos de outros lugares (estados, países). Em contrapartida, existem fatores que influenciam para escolha da loja física, como: entrega, segurança, contato direto com produto e com vendedor.

4.4 Decisão de Compra

Após obter todas as informações necessárias e avaliar as alternativas, eis que nessa etapa o consumidor irá efetuar a compra. Este processo se inicia quando o consumidor visita a loja (virtual ou física), escolhe o produto e efetua o pagamento. Durante esse procedimento, podem ocorrer vários fatores impeditivos para que este seja completado com sucesso. Blackwell[4], afirma que os consumidores podem mudar de opinião e comprar algo totalmente diferente do que era previsto, eles também podem mudá-la devido a ocasiões econômicas, promoções, falta do produto ou até mesmo por falta de motivação.

No Comércio Eletrônico, os fatores aumentam por causa das questões técnicas, como: tempo de resposta do site, interface, usabilidade, erros inesperados, velocidade da Internet, entre outros. Caso alguns dos problemas citados anteriormente ocorram, o usuário irá desistir da compra, ou irá procurar outra loja para finalizá-la. Isso, ocorre devido aos e-consumidores serem mais impacientes quando estão realizando suas compras no varejo virtual. Martins[15] cita que a falta de informações sobre produtos e um serviço de má qualidade podem fazer o consumidor desistir da compra.

Uma compra bem sucedida pode fazer com que o cliente se torne um “cliente fixo”, ou fidelizado, e esse é objetivo de toda organização. Quando este passa por uma compra, e fica satisfeito, o mesmo irá armazenar em sua memória determinada loja, isso faz com que, em necessidades futuras, o cliente volte a procurá-la. Caso aconteça o contrário, de encontrar muitos problemas, além de não comprar, irá reclamar de forma pública (amigos, familiares, Internet, etc), e isso pode trazer grandes prejuízos para a empresa.

4.5 Avaliação pós-compra

Costa[6] ressalta que um consumidor só ficará satisfeito/insatisfeito após ter consumido(usado) um produto/serviço que foi adquirido.

Muitas lojas virtuais possibilitam que os consumidores postem comentários e façam avaliação de um determinado produto que foi ou será adquirido. Principalmente após o

uso, e a experiência adquirida, o consumidor irá compartilhar sua opinião tanto no site em que efetuou a compra, quanto na rede de modo geral.

5. TRABALHOS RELACIONADOS

Fell, Rodrigues e Vieira[21] fazem uma avaliação da experiência de compra no *E-Commerce* a partir da taxonomia da usabilidade. Ainda segundo estes, a construção de uma loja virtual deve ser feita através da experiência do usuário, quando isso não ocorre, o consumidor encontrará dificuldades para finalizar uma compra. Oliveira[19] fala que na atualidade, com o crescimento do Comércio Eletrônico, é indispensável que as organizações deixem de conhecer o processo de decisão de compra através da Internet. Gonçalves[10] complementa dizendo que o processo de consumo dos e-consumidores é um item fundamental para elaboração de um *E-Commerce* de qualidade. Outros fatores que também devem ser considerados na elaboração do Comércio Eletrônico e que fazem parte da avaliação dos autores, Fell, Rodrigues e Vieira[21] são: consistência, feedback, metáforas, minimização de carga de memória, eficiência no diálogo, exibição exclusiva de informação relevante, uso de rótulos, abreviações e mensagens claras, mecanismo de ajuda, prevenção de erros e tratamento de erros.

Costa[6] realiza um estudo sobre o comportamento de consumidores de nível superior que realizam compras pela Internet. Já Miranda[17] estuda as variáveis comportamentais determinantes na decisão de compra nas relações virtuais entre usuários da Internet integrando os princípios de comportamento do consumidor. Em ambos os trabalhos, a comodidade apresenta-se em destaque como fator de escolha da Internet para efetuar a compra. Em contrapartida, a insegurança juntamente com a falta de contato com o produto são fatores que levam o consumidor a não escolher este meio para finalizá-la. Costa[6] apresenta também que a maioria dos compradores virtuais planeja ou sempre planejam suas compras e que 75,5% dos entrevistados levam a apresentação do site como critério na hora da escolha. E Gonçalves[10] completa está informação dizendo que “Forma, conteúdo e comportamento devem integrar-se, de maneira que o resultado final permita ao usuário utilizar as ferramentas disponíveis com consistência e facilidade”.

Migliorini[16] destaca que a falta de usabilidade implica na rejeição da Internet móvel. Gonçalves[10] em seu trabalho, desenvolveu um aplicativo, baseando-se em estudos feitos através da plataforma desktop do *E-Commerce*, somente desta forma foi possível perceber o que seria usado e o que seria ocultado na plataforma de *M-commerce*. O mesmo demonstra a importância do profissional de design que será o responsável por apresentar um layout agradável, focado na experiência do usuário, que facilita na hora da compra.

Ainda segundo Migliorini[16], a lentidão causada no carregamento das páginas, dificuldade de navegação e a inacessibilidade de alguns sites, são outros fatores que impedem o usuário de uma navegação agradável pelo celular. Estes itens decorrem dos mesmos estudados por Fell, Rodrigues e Vieira[21], que seriam: consistência e feedback.

Em seu trabalho, Oliveira[19] aborda o comportamento de consumidores de livros que utilizam a Internet para efetuar uma compra, separando pelas etapas de comportamento do consumidor propostas neste trabalho. De acordo com sua pesquisa, as fontes externas são as mais consultadas, e, in-

fluenciadoras no início do processo de compra.

A partir dos trabalhos apresentados anteriormente, é possível perceber algumas dificuldades encontradas no Comércio Eletrônico de modo geral, bem como alguns fatores que influenciam no processo de compra. No que se refere às plataformas *M-commerce* e *E-Commerce*, nenhum dos autores acima realizou uma comparação entre estas, mas é importante salientar que todos eles fazem uma abordagem referente à interface da loja virtual, e mostram que a usabilidade, a navegabilidade, o tempo de respostas, a disposição dos elementos na tela, podem influenciar no processo decisório em ambas as plataformas.

6. METODOLOGIA

Esta pesquisa tem como objetivo apresentar as características de comportamento do consumidor nas compras realizadas por dispositivos móveis e pelo computador, e, apresentar resultados obtidos por meio de questionários com questões objetivas.

A primeira etapa da pesquisa foi realizada através de um amplo levantamento bibliográfico a fim de especificar as características comportamentais do consumidor virtual. Já a segunda etapa foi definir quais dessas características e fatores comportamentais que deveriam ser comparados entre o *M-Commerce* e *E-commerce*.

A terceira etapa da pesquisa foi a elaboração do questionário, baseando-se, nos dispositivos que as pessoas utilizam para efetuar uma compra pela Internet, e, dividindo essas pessoas em três grupos: pessoas que só compram pelo computador/notebook, pessoas que só compram pelo celular/tablet, pessoas que compram pelo computador/notebook e pelo celular/tablet.

O método para coleta de dados escolhido foi a aplicação de um questionário, constituído por um total de 16 questões, sendo elas objetivas e do tipo múltipla escolha com uma única resposta e as do tipo matriz, que possui as colunas: Computador/Notebook e Celular/Tablet, de forma que, o usuário, pudesse responder de acordo com as compras realizadas em cada dispositivo. As linhas da matriz representam afirmações e características do processo de compra na Internet.

O questionário foi aplicado de forma online através do link <https://pt.surveymonkey.com/s/3TDH6S6>, usando o SurveyMonkey, que é um sistema bastante consolidado no mercado para aplicação de questionários online. O link foi compartilhado através de redes sociais, como: Facebook, WatSapp e LinkedIn. Não houve restrições quanto a sexo, faixa etária, profissão, sendo o questionário de forma livre para qualquer pessoa apto ou com conhecimento necessário a responder. O questionário poderia ser respondido através de computadores ou celulares com acesso a Internet a qualquer momento.

As perguntas estão divididas em 03 páginas, cada página com o objetivo de coletar hábitos do consumidor na Internet. A primeira página, com um total de 6 perguntas, procura identificar o perfil dos consumidores virtuais e se esses tem hábito de comprar pela Internet. Caso o consumidor respondesse que não, o questionário era encerrado.

Na segunda página, os entrevistados responderam questões do tipo, quais produtos/ serviços que são comprados/ contratados pela Internet e qual o motivo que os leva a desistir de uma compra na Internet.

E por último, a terceira página tem como objetivo, identificar o comportamento do consumidor no processo de com-

pra, nos dispositivos móveis e nos computadores. Essa página possui questões sobre a quantidade de produtos, valor das compras, segurança, forma de pagamento e uma comparação direta entre os dois dispositivos.

A seguir será apresentada a análise estática dos dados e serão apresentados os resultados.

7. RESULTADOS

Os resultados obtidos na pesquisa são apresentados, seguindo a ordem do questionário, e, com objetivo de apresentar resultados mais consistentes foi aplicado o relacionamento entre algumas respostas.

A análise dos dados foi realizada com base nas respostas de 163 questionados através da plataforma Survey Monkey, durante um período de 30 dias.

Na primeira parte do questionário foi possível identificar o perfil dos consumidores online, tais como: faixa etária, sexo, grau de escolaridade, renda familiar. Dentre esses prevaleceram: Homens (65%), idade entre 18 e 24 anos (47,2%), com nível superior incompleto (62,6%), e, seguindo o critério adotado pela Associação Brasileira de Empresas de Pesquisa (ABEP), em setembro de 2014, a classe social que prevaleceu na pesquisa foi a B1 (41,1%), que possui renda familiar entre R\$3.119,00 até R\$6.006,00.

Os resultados mostram também, que a maioria das pessoas possui tanto celular quanto computador com acesso à Internet (92,6%), e desses, a maioria tem hábito de comprar pela Internet (89,6%), como mostra a Tabela 01.

Tabela 1: Hábito de comprar pela Internet

Você tem o hábito de comprar pela Internet?		
Opções de Resposta	Porcentagem	Contagem
Sim	89,6%	146
Não	10,4%	17
Pessoas que responderam		163
Pessoas que não responderam		0

Na segunda parte tem início a comparação entre o processo de compra realizado pelo *E-Commerce* e pelo *M-Commerce*, e, apresenta também, o resultado da categoria de produtos e serviços que mais são comprados nas duas plataformas.

A tabela 2 mostra as 3 categorias de produtos que foram mais selecionadas em cada dispositivo. É importante ressaltar que a questão permitia mais de uma resposta.

Tabela 2: Categorias de Produtos Compradas

Computador\ Notebook			Celular\ Tablet		
Categoria	Quantidade de Respostas	%	Categoria	Quantidade de Respostas	%
Eletrônicos e Eletrodomésticos	110	82,71%	Artigos de Informática	19	14,29%
Artigos de Informática	93	69,92%	Livros e Revistas	15	11,28%
Vestuário	68	51,13%	Games	15	11,28%

É possível perceber com os resultados apresentados acima que os produtos comprados pelo computador são diferentes dos produtos comprados pelo celular, tendo em comum somente Artigos de Informática, mas em proporções bem menores. Outro fator de destaque é que a categoria de produtos mais comprados pelo computador (Eletrônicos e Eletrodomésticos (82,71%)), não aparece entre as três categorias mais selecionadas nas compras pelo celular, já que esta foi escolhida somente por 10,53% dos consumidores.

O mesmo acontece na contratação de serviços pela Internet, onde pelo computador a maioria dos consumidores con-

trata serviços de Download de Filmes (61,83%) e Compram ingressos (56,49%). E pelo celular destaque-se a contratação de serviços de Telefonia e Internet (25,95%) e Download de Músicas (23,66%), ou seja, os serviços contratados pelo celular são de, preferência, aqueles onde a pessoa pode utilizar no próprio aparelho.

A terceira parte do questionário apresenta resultados sobre o comportamento do consumidor no processo de compra, e, as tabelas 3 e 4 apresentam um relacionamento das questões sobre a quantidade de produtos comprados, o valor gastos nas compras e a forma de pagamento em cada um dos dispositivos.

Tabela 3: Dados Referentes a Última Compra Realizada no E-Commerce

Computador/Notebook					
Qnt. de Produtos	%	Valor gasto nas compras	%	Forma de pgto	%
2	26,77%	De R\$101,00 até R\$300,00	28,13%	Cartão Crédito	84,13%
1	25,98%	De R\$301,00 até R\$500,00	16,41%	Boleto	34,92%
3	20,47%	De R\$51,00 até R\$100,00	10,94%	Pag Seguro	18,25%

Tabela 4: Dados Referentes a Última Compra Realizada no M-Commerce

Celular/Smartphone					
Qnt. de Produtos	%	Valor gasto nas compras	%	Forma de pgto.	%
1	15,75%	De R\$10,00 até R\$50,00	9,38%	Cartão Crédito	31,75%
2	6,30%	De R\$101,00 até R\$300,00	7,03%	Boleto	7,14%
3	2,36%	De R\$301,00 até R\$500,00	5,47%	PayPal	7,14%

Como dito anteriormente os consumidores do M-Commerce preferem comprar produtos ou serviços que podem ser utilizados no próprio dispositivo móvel. Na Tabela 4 é possível concluir que esses consumidores compram na maioria das vezes apenas 1 produto/serviço de baixo valor, e, relacionando esse resultado com as categorias selecionadas, frequentemente essas compras são de aplicativos para o próprio celular ou produtos de valor baixo como livros e revistas.

A preferência pelo pagamento através do cartão de crédito ainda é unânime pelos consumidores dos dois dispositivos. Esse é o meio de pagamento mais atraente para estes, pois, podem parcelar suas compras, principalmente quando o valor é alto. As compras realizadas pelo cartão de crédito também proporcionam mais segurança por causa da entrega do produto, que pode chegar danificado, não atender as expectativas ou até mesmo nem ser entregue. Com isso o consumidor se resguarda de não pagar o produto logo de uma vez.

Outro aspecto identificado no comportamento do consumidor é em relação ao processo de pesquisa de produtos na Internet, onde o resultados da tabela 5 e 6 demonstram que

os consumidores pesquisam muito, antes de tomar a decisão sobre a compra, tanto no computador quanto nas compras pelo celular.

Tabela 5: Pesquisa do Consumidor no E-Commerce

Computador/Notebook	
Quantidade de sites visitados antes de efetuar a compra.	
Mais que 5	38,58%
3	29,92%
4	16,54%

Tabela 6: Pesquisa do Consumidor no M-Commerce

Celular/Tablet	
Quantidade de sites visitados antes de efetuar a compra.	
Mais que 5	18,90%
3	11,81%
2	7,87%

Atualmente os dois dispositivos permitem que o consumidor consulte com facilidade varias lojas virtuais, e, com auxílio de ferramentas de busca como: Buscapé, Zoom e BondFaro proporcionam mais praticidade na hora da pesquisa.

Ainda com relação ao processo de pesquisa, os dados das tabelas 7 e 8 mostram que os consumidores do E-Commerce e do M-Commerce desistem de uma compra na Internet mais por motivos financeiros e pelo fato de não encontrarem o produto que buscam no site. Além disso, é possível afirmar que os consumidores do M-Commerce também se preocupam com erros inesperados que aparecem na pesquisa e dificuldades de navegação. Por se tratar de uma interface menor, onde o consumidor tem o contato diretamente com na tela do aparelho, os compradores do mobile commerce são mais impacientes a erros não tratados, ou quando o programa é encerrado de forma inesperada.

Tabela 7: Motivos pelo Qual o Consumidor Desiste de uma Compra no E-Commerce

Computador/Notebook	
Motivos que levam consumidor a desistir da compra	
Quando o valor do frete é alto	79,55%
Quando o preço está mais alto que na loja física	78,03%
Quando não encontro os produtos que desejo	73,48%

Tabela 8: Motivos pelo Qual o Consumidor Desiste de uma Compra no M-Commerce

Celular/Tablet	
Motivos que levam consumidor a desistir da compra	
Quando não encontro os produtos que desejo	40,15%
Quando o preço está mais alto que na loja física	33,78%
Quando ocorre um erro inesperado	37,12%

Até o momento é possível perceber algumas diferenças no perfil dos consumidores em relação às compras realizadas pela Internet, como, tipo produtos/serviços comprados, valor gasto nas compras, formas de pagamento e motivos que levam estes a desistirem de uma compra. Neste ponto, é possível perceber que os consumidores estão sujeitos a mudança de comportamento de acordo com o dispositivo em

que está efetuando a compra. No entanto a tabela 9 mostra um resultado que compara a utilização dos dispositivos (computador e celular) durante o processo de compra online.

Tabela 9: Processo de Compra na Internet

Atividades realizadas no processo de compra online de acordo com cada dispositivo		
	Computador /Notebook	Celular /Tablet
Visito sites em busca de promoções de produtos e serviços	79,69%	32,03%
Faço cadastro para comprar no site	82,81%	25,00%
Utilizo o chat para falar com vendedor	30,47%	5,47%
Visualizo comentários de outros compradores	83,33%	31,06%
Visualizo as características do produto	88,28%	35,94%
Adiciono produtos ao carrinho de compra	83,59%	26,56%
Seleciono a forma de pagamento	88,28%	25,78%
Efetuo o pagamento	88,28%	26,56%
Faço Acompanhamento do Pedido	86,72%	33,59%
Faço um comentário ou avaliação do Produto	37,50%	14,84%

De acordo com a tabela anterior é possível perceber a fragilidade do *M-Commerce* com relação às vendas realizadas pela Internet, pois, todas as atividades são realizadas mais no computador do que no celular. Entretanto, é importante ressaltar que as atividades que são mais executadas no celular são as de pesquisa e visualização dos produtos e acompanhamento do pedido. Esse fato é justificado pelo uso do celular no dia a dia das pessoas, onde a possibilidade de acesso à Internet de qualquer lugar proporciona essa comodidade para que os consumidores com tempo livre busquem produtos e promoções na Internet e possam acompanhar o pedido caso já tenha efetuado a compra.

A tabela 10 exhibe os motivos pelo qual os consumidores virtuais ainda preferem mais efetuar a compra pelo computador do que pelo celular.

Tabela 10: Motivos que Influenciam o Consumidor na Escolha de cada Dispositivo

Atividades realizadas no processo de compra online de acordo com cada dispositivo		
	Computador /Notebook	Celular /Tablet
Tenho mais habilidade para efetuar a compra	86,72%	3,13%
Tenho mais segurança para efetuar a compra	81,25%	5,47%
Tenho mais comodidade para efetuar a compra	68,75%	24,22%
Utilizo mais para pesquisar ofertas na Internet	55,47%	34,38%
Utilizo mais para comparar preços na Internet	67,97%	21,88%
Utilizo mais durante o dia para navegar na Internet/rede social	22,66%	65,63%

Como falado anteriormente, o uso do celular no dia a dia das pessoas está maior do que a utilização de computadores, entretanto as compras ainda continuam maiores no computador por motivos de segurança (81,25%) e mais habilidade (86,72%) para conseguir efetuar a compra.

A segurança é um grande problema enfrentado nas compras pela Internet, e, muito mais no *M-Commerce*. Os consumidores se preocupam ao fornecerem dados pessoais e efetuarem pagamento, pois, não sabem se de fato esses serão mantidos em segurança. Baseando-se nisso, existem algumas formas do consumidor conseguir se proteger ao efetuar uma compra na Internet, que será visto na tabela 11.

Tabela 11: Segurança nas Compras Online

Cuidados que consumidores tem nas compras online de acordo com cada dispositivo		
	Computador /Notebook	Celular /Tablet
Compro somente em sites que possui certificados de segurança	83,72%	27,91%
Possuo antivírus instalado	66,67%	17,05%
Sempre me conecto em uma rede segura para realizar a compra	64,34%	18,60%
Recebo mensagens/e-mails informando sobre dados de pagamento	79,84%	31,78%
Já tive dados roubados/ alterados e usados por terceiros	5,43%	1,55%

Os consumidores se preocupam na hora da compra, se o site é seguro, entretanto muitos se conectam em redes abertas o que é um descuido, já que, em uma rede aberta também é possível que os dados sejam roubados. Para o celulares o risco é maior já que este dispositivo é mais propício a roubo e os consumidores não se preocupam tanto a segurança, já que poucas pessoas possuem antivírus instalado e na maioria das vezes estão conectados a redes abertas.

8. CONCLUSÕES

O *M-Commerce* vem aumentando sua porcentagem de vendas a cada ano e o número de pessoas que se conectam a Internet através de dispositivos móveis supera as expectativas do mercado. As previsões são de que os números aumentem nos próximos anos.

Esta pesquisa apresenta o comportamento do consumidor no processo de compra pela Internet nos dispositivos móveis e nos computadores, o perfil dos consumidores, os hábitos e expõe dados sobre o mercado brasileiro de Comércio Eletrônico. Mostra também que as pessoas utilizam mais os dispositivos móveis durante o dia (65,63%), entretanto, quando necessitam realizar uma compra na Internet preferem efetuar a compra no computador. É importante destacar que todo o processo de compra realizado diretamente no computador, desde a busca pelos produtos até a finalização da compra sendo que os dispositivos móveis apenas auxiliam neste processo, como nas atividades de pesquisar os produtos e para acompanhamento do pedido.

A sociedade está passando por uma mudança cultural, que já havia ocorrido antes com o surgimento do *E-Commerce*, onde os consumidores foram deixando as compras na loja física e migrando cada vez para a Internet, o mesmo deve acontecer com as compras realizadas através dos *smartphones*. De acordo com a pesquisa 92,6% dos entrevistados possui tanto celular quanto computador com acesso à Internet e 89,6% tem hábito de comprar pela Internet.

Os consumidores estão se adaptando a comprar neste tipo de dispositivo, mas é necessário que as operadoras de tele-

fonia melhorem sua infraestrutura de rede para uma melhor conexão do consumidor e que as empresas responsáveis pelo desenvolvimento dos sites de *M-commerce* melhorem as interfaces de modo a proporcionarem uma melhor experiência para usuário, pois 86,72% dos consumidores possuem mais habilidade para efetuarem a compra pelo computador.

É necessário também que o consumidor se sinta seguro ao efetuar compras no *M-Commerce*, já que 81,25% dos consumidores dizem ter mais segurança ao efetuarem uma compra no computador e a partir disso os consumidores passariam a consumir produtos com valores mais altos já que o valor gasto em média nas compras pelo celular é de R\$ 10,00 a R\$ 50,00 e de apenas um produto (15,75%). Outro problema que também deve ser resolvido é em relação aos erros que ocorrem durante a compra, já que 37,12% das pessoas desistem da compra quando ocorre um erro inesperado. As redes de varejo necessitam criar campanhas promocionais para atrair o consumidor para comprar pelo *smartphone*, pois de acordo com 33,78% dos consumidores de *M-Commerce* que responderam a pesquisa, desistem da compra quando o preço está maior que na loja física.

Apesar do abrangente campo de estudos no que refere a compra pela Internet, uma oportunidade seria explorar o desenvolvimento de interfaces de vendas para dispositivos móveis, de forma a proporcionar mais facilidade na hora da compra. Outra opção de estudo diz respeito melhores estratégias de Marketing de forma a fomentar as vendas através de dispositivos móveis.

9. REFERÊNCIAS

- [1] F. ALEXANDRINI. *Perfil empresarial na prática do e-commerce: comercialização eletrônica*. Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia de Produção., 2000.
- [2] C. ANGELO and J. SILVEIRA. *Varejo Competitivo*. Editora Atlas, 3th edition, 1999.
- [3] C. ANGELO and J. SILVEIRA. *Varejo Competitivo*. Editora Atlas, 6th edition, 2001.
- [4] R. BLACKWELL, P. MINIARD, and J. ENGEL. *Comportamento do Consumidor*. Editora Thomson Pioneira, 9th edition, 2005.
- [5] R. CARVALHO. Negócios Eletrônicos na modalidade B2C – Business to Consumer - Um estudo de caso de uma empresa varejista de produtos eletroeletrônicos em Fortaleza, 2010. Monografia (Bacharel Ciência da Computação) - FLF (Faculdade Lourenço Filho Ciência da Computação), Fortaleza, Brasil.
- [6] F. COSTA. Comércio Eletrônico: Hábitos do Consumidor na Internet, 2009. Dissertação (Mestrado Profissional em Administração) - FPL (Fundação Pedro Leopoldo), Pedro Leopoldo, Brasil.
- [7] H. DEITEL, P. J. DEITEL, and K. STEINBUHLER. *E-business e e-commerce para administradores*. Pearson Education, 2004.
- [8] E-BIT. Comércio eletrônico no brasil. <http://www.ebit.com.br/webshoppers>, 2013. Último Acesso em: 07-03-2014.
- [9] J. ENGEL, R. BLACKWELL, and P. MINIARD. *Comportamento do Consumidor*. LTC Editora, 8th edition, 2000.
- [10] F. GONÇALVES. M-Commerce: Projetando o Processo de Compra para Smartphones, 2013. Monografia (Bacharel Design Digital) - UFPEL (Universidade Federal de Pelotas), Pelotas, Brasil.
- [11] HI-MIDIA. E-commerce: Comportamento dos e-consumidores. <http://www.hi-midia.com/pesquisas/maio2012-e-commerce-comportamento-dos-e-consumidores/>, 2012. Último Acesso em: 18-03-2014.
- [12] HI-MIDIA. Mercado de mobile e de m-commerce no brasil. <http://www.hi-midia.com/pesquisas/outubro-2012-pesquisa-analisa-o-mercado-de-mobile-e-de-m-commerce-no-brasil>, 2012. Último Acesso em: 07-03-2014.
- [13] P. KOTLER and K. L. KELLER. *Administração de Marketing*. Pearson Prentice Hall, 12th edition, 2006.
- [14] S. MARTI. Mobile commerce não é tendência, é um fato atual. <http://www.ecommercebrasil.com.br/artigos/mobile-commerce-nao-e-tendencia-e-um-fato-atual>, 2014. Último Acesso em: 18-03-2014.
- [15] F. MARTINS. 4 cuidados para não perder clientes virtuais. <http://ecommercenews.com.br/artigos/tutoriais/4-cuidados-para-nao-perder-clientes-virtuais>, 2012. Último Acesso em: 16-03-2014.
- [16] A. MIGLIORINI. Mobile Commerce – E-commerce, 2009. Trabalho de curso (Bacharelado em Administração) - UNIVEM (Centro Universitário Eurípides de Marília), Marília, Brasil.
- [17] C. MIRANDA and D. ARRUDA. E-produtos e variáveis comportamentais determinantes de compra no varejo virtual: Um estudo com consumidores brasileiros. *REAd*, 10(1), 2004.
- [18] J. O'BRIEN. *Sistemas de Informação e as Decisões Gerenciais na Era da Internet*. Saraiva, 2004.
- [19] E. OLIVEIRA. Comportamento do consumidor - processo de decisão de compra de livros pela internet, 2007. Dissertação (Mestrado Profissional em Administração) - USP (Universidade de São Paulo), São Paulo, Brasil.
- [20] O. M. Planet. Nosso planeta mobile: Brasil - como entender o usuário de celular. <http://services.google.com/fh/files/misc/omp-2013-br-local.pdf>, 2013. Último Acesso em: 28-03-2014.
- [21] G. RODRIGUES, A. VIEIRA, and A. FELL. Avaliação de experiência de compra no e-commerce a partir da taxonomia da usabilidade. *Encontro de Estudos Sobre Tecnologia, Ciência e Gestão da Informação*, 2013.
- [22] K. SILVA, R. PRADO, M. SILVA, and J. SILVA. Processo de decisão de compra: Um estudo sobre o comportamento do consumidor virtual. *XXXIII Encontro Nacional de Engenharia de Produção*, 2013.
- [23] TELEBRASIL. Número de acessos em banda larga chega a 137 milhões. <http://www.telebrasil.org.br/sala-de-imprensa/releases/5490-numero-de-acessos-em-banda-larga-chega-a-137-milhoes-em-janeiro>, 2013. Último Acesso em: 07-03-2014.

Reconhecimento de árvores usando visão computacional para prevenir interrupções no sistema de distribuição de energia elétrica.

Alternative Title: Recognition of trees using Computer Vision to prevent interruptions in the Power Distribution System

Heuber G. F. de Lima
Universidade Federal de Goiás
Instituto de Informática
hebergustavo@inf.ufg.br

Ronaldo M. da Costa
Universidade Federal de Goiás
Instituto de Informática
ronaldocosta@inf.ufg.br

Anderson da S. Soares
Universidade Federal de Goiás
Instituto de Informática
anderson@inf.ufg.br

Gustavo T. Laureano
Universidade Federal de Goiás
Instituto de Informática
gustavo@inf.ufg.br

RESUMO

A distribuição de energia elétrica é uma atividade fundamental, já que muitas pessoas dependem deste serviço. Falhas no sistema de distribuição ocorrem a partir de vários fatores que podem danificar o sistema e, portanto, interromper o fornecimento de energia elétrica. Dentre os vários fatores que podem causar problemas este trabalho propõe a detecção automática de árvores perto ou mesmo sob na rede de distribuição. A fim de evitar que as árvores possam forçar ou mesmo causar a ruptura dos cabos de distribuição, são feitas as podas das árvores que apresentam algum tipo de risco para a rede. No entanto, essa atividade é geralmente o manual fazendo com que as equipes necessitem vasculhar toda a rede para identificar os problemas. O principal objetivo deste trabalho é propor um processo, baseado em visão computacional, que permite a identificação automática de árvores próximas ou sob a rede de distribuição de energia a partir de imagens aéreas fornecidas pelo Google Earth.

Palavras-Chave

Reconhecimento de Padrões, Visão Computacional, Identificação de Árvores.

ABSTRACT

The electrical distribution is a critical activity since many people depend of this service. Faults in the distribution system occur from several factors that can damage the system and therefore interrupt the supply of energy. Among the various factors that may cause problems this work proposes a automatic detection of trees near or even in the distribution network. In order to avoid that the trees to force or even rupture of the distribution cables, are made the pruning of the trees that have some kind of risk to the network. However, this activity is usually manual and teams must sift through all the network for problems. The main objective of this work is to propose a process, based on computer vision, which allows the automated identification of nearby trees or under the power distribution network from aerial images provided by Google Earth.

Categories and Subject Descriptors

I.5.4 Applications: Pattern Recognition.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.

Copyright SBC 2015.

General Terms

Algorithms, design.

Keywords

Pattern recognition, computer vision, tree identification.

1. INTRODUÇÃO

Geração, Transmissão e Distribuição são conceitos diferentes no âmbito das concessionárias de energia. Os cabos que estão presentes nas ruas das cidades, que prosseguem por via aérea ou subterrânea, transformadores, postes, cabos, religadores entre outros equipamentos fazem parte da chamada Rede de Distribuição de Energia[1].

Vários fatores podem gerar problemas que afetam a distribuição de energia para os consumidores finais, tais como: sobrecarga do sistema em função de um elevado consumo em um determinado momento, curto circuito na rede de distribuição em função de equipamentos sem a devida manutenção, árvores que forçam a rede fazendo com que os cabos se toquem ocasionando curto circuito, ou até mesmo rompimento dos cabos em função de agentes externos, tais como árvores, caminhões e etc.[2]

Em relação aos equipamentos obsoletos, as concessionárias de energia tem preparado rotinas de manutenção programada onde considera-se o tempo de vida desses equipamentos [2][3] ou ainda a adoção de equipamentos chamados Religadores Automáticos[4], que tentam religar a energia de forma automática ao detectarem a interrupção[4]. Em relação ao consumo excessivo, uma prática adotada por alguns países, é a adoção do chamado “horário de verão” onde a hora nacional é adiantada em 1 hora, fazendo com que o horário de pico, que é no final da tarde, ainda possua luz natural diminuindo o consumo de energia no horário em função da diminuição do uso de luzes artificiais. Agora, a relação entre as árvores plantadas nas cidades com a rede de distribuição de energia para os consumidores talvez seja um dos fatores que mais geram problemas na distribuição. As falhas na realização de podas preventivas de árvores que interferem na rede de baixa tensão contribuem para as frequentes quedas no fornecimento de energia.

Na maioria das cidades as árvores, independente do seu porte, de uma forma geral estão posicionadas nas calçadas, mesmo lugar em que está a rede elétrica para a distribuição. Na estação das chuvas a força dos ventos ou até mesmo a força que a chuva exerce sobre a árvore faz com que elas se inclinem sobre a rede, fazendo com que os cabos se encontrem gerando um curto circuito na rede e, por consequência, a interrupção do fornecimento; sem falar na possibilidade da queda dessa árvore

sobre a rede de energia fazendo com que os cabos sejam rompidos, o que acarreta um tempo maior de interrupção do fornecimento uma vez que a rede de distribuição precisaria ser reparada [2][3][5].

A contenção para esse tipo de problema é feita através da poda das árvores que estão junto à rede de distribuição. Em geral, o serviço de poda se dá através de cortes programados e através da reclamação do consumidor. Sobre a poda programada, ao iniciar a estação das chuvas é promovida a podas das árvores que atrapalham a rede através de um mutirão com essa finalidade. Entretanto, esse trabalho é todo manual, ou seja, as equipes precisam percorrer a rede em busca de possíveis problemas, o que gera um gasto excessivo de tempo, recursos e de material humano. Sobre a poda por reclamação, na maioria das vezes o consumidor reclama de forma reativa; ou seja, quando o problema já ocorreu e por consequência o fornecimento de energia já ter sido prejudicado [1][2][3][5].

As agências de controle da qualidade e fornecimento da energia elétrica, de certa forma costumam estabelecer compensações para os consumidores que tiveram o fornecimento de energia elétrico prejudicado. Algumas estabelecem ainda limites para esses indicadores e quando há violação desses limites, a distribuidora deve compensar financeiramente a unidade consumidora. Enfim, não é do interesse de nenhuma distribuidora de energia ter o fornecimento de energia interrompido ocasionado por árvores tocando ou caindo sobre a sua rede de distribuição [6][7][8].

Um método de identificação automática de árvores seria de grande auxílio na identificação e antecipação de possíveis problemas com árvores e a rede de distribuição de energia. Mas um sistema dessa natureza não se restringiria à essa finalidade. Um outro exemplo seria a utilização desse tipo de ferramenta na segurança pública, para prevenção de acidentes ou no planejamento e controle da vegetação em grandes centros urbanos[5]. Na agricultura, este tipo de aplicação pode ser utilizado para dimensionamento e controle de culturas como plantações de laranja, manga e etc. [9][10][11][12][13].

Em outros trabalhos correlatos as propostas demandam recursos onerosos, como em Othman [14] em que a proposta é identificar as árvores a partir de um Scanner Terrestre Laser (TLS em Inglês), equipamento que tem um custo elevado. Já Backes, [15], Idrissa [16], Jain [17], Lee [18] e Wang [19] propõem o reconhecimento de árvores por meio da identificação da textura ou formato das folhas, e utiliza-se de imagens em alta resolução que possuem um alto custo de aquisição. A proposta deste trabalho é utilizar recursos que estão gratuitamente disponíveis, que é o caso do Google Earth.

A proposta desse trabalho é apresentar um processo de identificação automática de árvores que potencialmente prejudicariam a rede de distribuição de energia através do processamento de imagens captadas de aplicações como o Google Earth e utilização de algoritmos de visão computacional. Dessa forma com a adoção desse processo as concessionárias de energia poderão otimizar os seus recursos na poda programada das árvores possibilitando que as equipes de poda possam ser melhor programadas e destacadas para tal fim.

Este trabalho é composto de quatro seções: Inicialmente trata-se de uma introdução, na seção seguinte é apresentada uma proposta de arranjo para a detecção das árvores conforme o objetivo do

trabalho. Em seguida é apresentado o resultado do processo proposto e finalmente as conclusões obtidas.

2. PROPOSTA PARA DETECÇÃO AUTOMÁTICA DE ÁRVORES ATRAVÉS DE IMAGENS AÉREAS

Diante de todo o arcabouço tecnológico, a proposta para a identificação de árvores a partir de uma imagem obtida é demonstrado na Figura 1. A proposta apresentada é um modelo baseado na adoção de diversas metodologias de diferentes áreas e níveis de atuação, com o objetivo de criar um modelo estruturado e que consiga reunir as várias vantagens de cada metodologia escolhida, integrando-as. De acordo com o arcabouço do estado da arte em identificação de árvores através de imagens [20] [21][22][23][24][25][26] esses seriam os passos para capturar, processar e identificar corretamente uma árvore em uma imagem digital.

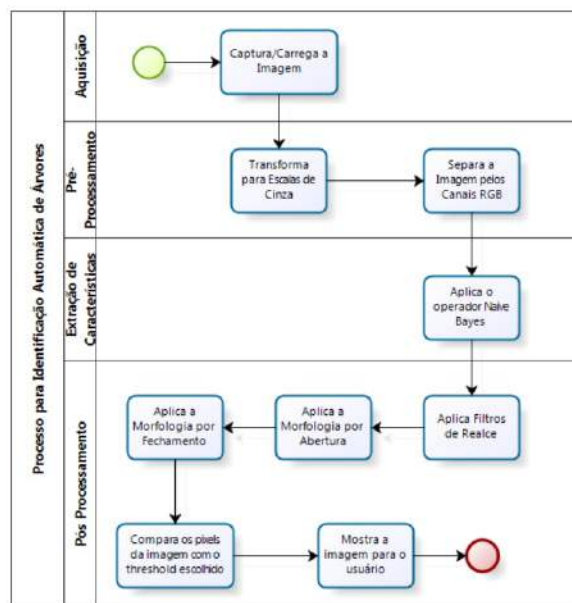


Figura 1. Proposta de Arranjo para a Identificação Automática de Árvores.

Passo 1 – Captura/Carrega a Imagem – Capturar as imagens a partir de uma aplicação que visualize a imagem de satélite do Google Earth e permita ao usuário acionar o verificador automatizado a fim de identificar as árvores daquela região. O processo de aquisição das imagens foi dividido em duas partes: Treinamento e Processamento. Para o treinamento, a partir de imagens aéreas que possuem árvores bem definidas, foi feito o recorte dessas árvores em uma determinada região a fim de identificar os parâmetros que classificam corretamente uma árvore. Um exemplo desse tipo de recorte pode ser visto na figura 2a. Posteriormente foram separadas imagens para o processamento: que são imagens aéreas, de zonas urbanas em que a intenção é identificar automaticamente as árvores presentes nas imagens. Para efeito de testes laboratoriais foram escolhidas, aleatoriamente, imagens de regiões urbanas que continham árvores a serem identificadas. Um exemplo de imagem para o processamento pode ser visto na figura 2b.

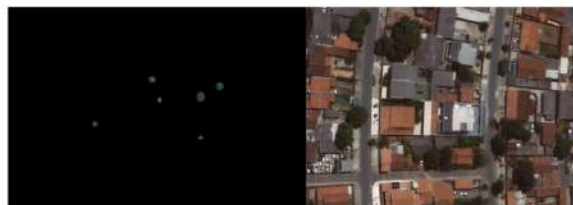


Figura 2. a) Exemplo de imagens de recorte para o treinamento b) Imagem para o Processamento

Passo 2 – Transforma para escalas de cinza – Passo fundamental para se processar imagens digitais é transformá-la em uma escala de tons de cinza [9][11]. A imagem digital em nível de cinza obtida a partir da imagem original é uma imagem na qual o valor de cada pixel é uma única amostra de um espaço de cores. Imagens desse tipo são tipicamente compostas com tons de cinza, variando entre o preto como a menor intensidade e o branco como maior intensidade [27][28].

Passo 3 – Separa a imagens pelos canais RGB – O objetivo é identificar a quantidade de cores Red – Vermelho, Green – Verde e Blue – Azul, presentes em cada um dos pixels da imagem. A imagem de entrada, obtida no passo 1, possui três bandas, sendo que cada uma delas representa uma das componentes do seu modo de cor. No caso deste trabalho a imagem, que era do modo RGB, seria comparada, pixel a pixel, com os níveis de cores obtidos no treinamento da imagem de recorte. De acordo com Casanova [20], esse método, obviamente, combinado com outros procedimento podem levar a um nível de acerto de 96% nas áreas analisadas.

Passo 4 – Aplica o Operador Naive Bayes – O objetivo desse passo é identificar a função que normaliza os níveis de cada uma das cores. Para tanto é necessário achar os valores médios e o desvio padrão de cada uma das cores. O resultado do processamento de uma determinada imagem, aplicado o operador de Naive Bayes pode ser acompanhado com a geração de um gráfico das Gaussianas, em que cada curva representa as cores RGB (red, green e blue).

Passo 5 – Aplicar Filtros de Realce – As técnicas de filtragem são transformações da imagem pixel a pixel, que não dependem apenas do nível de cinza de um determinado pixel, mas também do valor dos níveis de cinza dos pixels vizinhos. O processo de filtragem é feito utilizando matrizes denominadas máscaras, as quais são aplicadas sobre a imagem. Os filtros de Realce utilizam máscaras apropriadas ao realce de características de imagens obtidas por um sensor específico [29].

Passo 6 – Aplica a Morfologia por Abertura – A abertura em geral suaviza o contorno de uma imagem, quebra estreitos e elimina proeminências delgadas, a operação de abertura é usada também para remover ruídos da imagem [30][31][29].

Passo 7 – Aplica a Morfologia por fechamento – O fechamento funde pequenas quebras e alarga golfos estreitos elimina pequenos orifícios, se uma abertura cria pequenos vazios na imagem, um fechamento irá preencher ou fechar os vazios, estas operações podem remover muitos dos pixels brancos com ruídos, ou seja basicamente ele é igual a abertura só que primeiramente é feita a dilatação e, logo em seguida, é feita a erosão [30][31][29]. Em outras palavras o fechamento trabalha de um modo oposto ao método abertura, em que ela remove todos os pixels onde o ajuste do elemento estruturante não está dentro da imagem (primeiro plano), o fechamento enche todos os lugares onde o elemento estruturante não iria ajustar na imagem (plano fundo) [27].

Entretanto operações inversas, abertura e fechamento não irão estabelecer a imagem original.

Passo 8 – Compara os pixels da imagem com o threshold escolhido – O objetivo desse passo é identificar, com base no threshold escolhido, qual pixel estaria dentro do range de valores possíveis que foram processados anteriormente [27][32]. O threshold é gerado por um gerador automático de hipóteses que consideram que um determinado pixel pode ou não ser uma árvore.

Passo 9 – Mostra a imagem para o usuário – É o passo final. Mostra a imagem processada para o usuário.

3. RESULTADO DA PROPOSTA

De acordo com a proposta apresentada na seção anterior, foram feitos vários testes e processamentos para verificar o nível de acurácia do arranjo proposto. A princípio foram selecionadas 50 imagens das mais variadas regiões urbanas, e para cada região escolhida foi selecionado um recorte para o treinamento do algoritmo. Para cada imagem selecionada, contou-se manualmente as árvores presentes. Para efeitos de contagem, as árvores que aparentemente parecem ser uma só, ou seja, que possuem as copas ligadas, foram contabilizadas somente uma vez. Em seguida foi feito o processamento e apresentado o resultado final mostrando as árvores marcadas na imagem. Diante do resultado apresentado, foi refeita a contagem das árvores marcadas e contabilizadas a fim de se identificar as possíveis discrepâncias em relação à contagem inicial. A seguir, é mostrado o resultado do processamento de 10 imagens da cidade de Goiânia, capital do estado de Goiás, Brasil.

Tabela 1. Resultado do Processamento

Imagem	Árvores na imagem	Árvores Detectadas	Falsos Positivos	Verdadeiros Positivos	Verdadeiros Negativos
1	22	22	3	22	0
2	15	13	2	13	2
3	19	18	4	18	1
4	18	17	4	17	1
5	45	42	3	42	3
6	32	30	4	30	2
7	42	35	2	35	7
8	18	18	10	18	0
9	33	30	3	30	3
10	14	14	0	14	0

Na figura 3 é possível observar um exemplo de imagem de bairro predominantemente residencial, com cores bem definidas, com imagens de boa qualidade e com poucas sombras. Nestes casos o algoritmo conseguiu identificar as árvores com baixo nível de ruído, ou seja, com poucos falsos positivos. As árvores estão destacadas na imagem produzida ao final do processamento.

O método proposto ainda não distingue onde a árvore estaria localizada. O interessante seria que o método identificasse somente as árvores que estivessem sob a rede de distribuição, localizadas nas calçadas dos bairros. Esse nível de inteligência poderá ser atingido integrando o método proposto com a base de dados georreferenciada das Central Elétrica de Goiás – CELG, Brasil, objeto do estudo. Uma outra inteligência importante para o sistema seria a de identificar somente as árvores potencialmente prejudiciais à rede, ou seja, as que poderiam causar curtos ou rompimento da rede de distribuição. Para esse fim será necessário,

além de identificar o posicionamento das árvores sob a rede de distribuição, mas também identificar a altura e o tamanho das copas das árvores.



Figura 3. (a) Imagem Original e (b) Final do processamento

4. CONCLUSÃO

Como conclusão, entende-se que o processo para a identificação de árvores a partir de imagens obtidas no *Google Earth*, versão gratuita, é algo complexo uma vez que não existe uma padronização da qualidade da imagem. Outro fator, que é intrínseco ao problema, é a quantidade de informações presentes nas imagens: prédios altos que geram muitas sombras, casas com telhados das mais variadas cores, lotes vagos, arbustos, entre outros, são os fatores que dificultam o processamento e a identificação das árvores nas imagens. Existe ainda o fato das funções utilizadas precisarem ser calibradas com os mais variados parâmetros. Não foram encontrados na literatura exemplos de projetos que demonstrassem essa calibração. A calibração das funções foram obtidas através de experimentos de laboratório. Outro fator complicador no sucesso da identificação automática está na ordem de execução das funções de tratamento das imagens. A ordem das funções também influencia, em muito, o resultado final.

Enfim, a identificação automática das árvores, mesmo sendo algo complexo, o processo proposto consegue identificar as árvores dentro do contexto da imagem selecionada, observados os testes executados. Ou seja, dentro dos exemplos testados em laboratório, com parâmetros, comportamentos e limites definidos os resultados foram bastante satisfatórios. Entretanto, em função dos níveis de ruído encontrados em alguns dos experimentos, seria interessante partir para soluções em que a plataforma a ser utilizada, por exemplo, seja a versão paga do *Google Earth*. A versão paga, além da qualidade da imagem ser superior ao da versão gratuita, ela possui um intervalo de atualização reduzido, ou seja, as imagens são mais recentes, o que facilitaria, por exemplo, a identificação do estado da copa de uma árvore.

Para os trabalhos futuros é importante identificar as árvores que estão posicionadas especificamente sob a rede de distribuição, bem como marcar as árvores mais propensas a causar problemas na rede.

5. REFERÊNCIAS

- [1] Kagan, N., de Oliveira, C.C.C.B. & Robba, E.J., 2005. Introdução aos sistemas de distribuição de energia elétrica, Edgard Blücher.
- [2] Projeto SCOM 02, Comitê de Distribuição, "Manutenção de Sistemas de Distribuição", CELG – Centrais Elétricas de Goiás, 1988.
- [3] Projeto SCOM 08, Comitê de Distribuição, "Critérios para a Manutenção de Rede de Distribuição", CELG – Centrais Elétricas de Goiás, 1988.
- [4] NTC20, Setor de Normatização Técnica, "Religadores Automáticos", CELG – Centrais Elétricas de Goiás.
- [5] Projeto SCOM 20, Comitê de Distribuição, "Rotinas Básicas para a Manutenção de Rede de Distribuição", CELG – Centrais Elétricas de Goiás, 1988.
- [6] ANEEL, 2013, "Monitoramento da Qualidade – DEC e FEC", Superintendência de Fiscalização e Serviços de Eletricidade, 2013.

- [7] ANEEL, 2014, "Indicadores de Fornecimento ANEEL", Agência Nacional de Energia Elétrica, 2014.
- [8] ANEEL, 2011, "Resolução Normativa No 469", Agência Nacional de Energia Elétrica, 2011.
- [9] Brandtberg, T. Walter, F., "Automated delineation of individual tree crowns in high spatial resolution aerial images by multiple-scale analysis," *Mach. Vis. Appl.*, vol. 11, no. 2, Oct. 1998.
- [10] Gougeon, F. A., "A crown-falling approach to the automatic delineation of individual tree crowns in high spatial resolution aerial images," *Can. J. Remote Sens.*, vol. 21, pp. 274–284, 1995.
- [11] Gougeon, F. A. and Leckie, D. G., "The individual tree crown approach applied to Ikonos images of a coniferous plantation area," *Photogramm. Eng. Remote Sens.*, vol. 72, Nov. 2006.
- [12] Ke, Y. and Quackenbush, L. J., "Forest species classification and tree crown delineation using QuickBird imagery," in *Proc. ASPRS Annu. Conf.*, Tampa, FL, 2007.
- [13] Pinz, A. J., "A computer vision system for recognition of trees in aerial photographs," in *Proc. Int. Assoc. Pattern Recognit.*, 1991.
- [14] Othman, A., Voona, L. F. C. L. Y., Stolza, C. e Piboule, A. 2013, "Single tree species classification from Terrestrial Laser Scanning data for forest inventory", 2013.
- [15] Backes, A. R. and Bruno, O. M., "A new approach to estimate fractal dimension of texture images," In *International Conference on Image and Signal Processing*, 2008.
- [16] Idrissa, M. and Acheroy, M., "Texture classification using gabor filters. *Pattern Recognition Letters*", 2002.
- [17] Jain, A. K., Farrokhnia, F., "Unsupervised texture segmentation using gabor filters. *Pattern Recognition*, 1991.
- [18] Lee, C. L. and Chen, S. Y. "Classification of leaf images. *International Journal of Imaging Systems and Technology*", 2006.
- [19] Wang, Z., Chi, Z., and Feng, D. D., "Shape based leaf image retrieval", *IEEE Proceedings on Vision Image and Signal Processing*, 2003.
- [20] Casanova D., "Identificação de Espécies Vegetais por meio da análise da textura foliar", 2008.
- [21] Galloway, M. M., "Texture analysis using gray level run lengths. *Computer graphics and image processing*, Elsevier", v. 4, n. 2, 1975.
- [22] Haralick, R. M., Shanmugam, K., Dinstein, I. H., "Textural features for image classification. *Systems, Man and Cybernetics*", *IEEE Transactions on*, IEEE, n. 6, 1973.
- [23] Jensen, J. R. *Introductory digital image processing: a remote sensing perspective*. New Jersey: Prentice-Hall. 1986.
- [24] Neves, L. A. P., Neto, H. V. e Gonzaga, A.; *Avanços em Visão Computacional*, Ed. Omnipax, 2012.
- [25] Otsu, N., "A threshold selection method from gray level histograms", *IEEE Transactions on Systems, Man and Cybernetics*, 1979.
- [26] Tzionas, P., Papadakis, S., and Manolakis, D., "Plant leaves classification based on morphological features and a fuzzy surface selection technique". In *5th International Conference on Technology and Automation*, 2005.
- [27] Chen, C. H., Pau, L. F. and Wang, P. S. P., "Handbook of Pattern Recognition and Computer Vision." World Scientific, 1992.
- [28] Paes, F. C., Pinto, S. A. F., Dutra, F. V., "Um estudo exploratório para obtenção de índices de qualidade de vida urbana (iqv) utilizando imagens de alta resolução do Satélite Ikonos II", 2006.
- [29] Gonzalez, R. C., Woods R. E., Eddins S. L.; *Digital Image Processing Using Matlab*, Pearson Education, 2010.
- [30] Filho, O. M., *Processamento Digital de Imagens*, 3ª Ed., Pearson Education, 2011.
- [31] Gonzalez, R.C., Woods, R.E., *Processamento de Imagens Digitais*. Ed. Edgard Blücher, 2000.
- [32] Fukunaga, K., "Introduction to Statistical Pattern Recognition", Academic Press, 1990.

Um Sistema de Predição de Relacionamentos em Redes Sociais

Alternative title: A Link Prediction System in Social Networks

Luciano Antonio
Digiampietri
University of Sao Paulo
Av. Arlindo Bettio, 1000
Sao Paulo, SP, Brazil
digiampietri@usp.br

Caio Rafael do
Nascimento Santiago
University of Sao Paulo
Av. Arlindo Bettio, 1000
Sao Paulo, SP, Brazil
caio.rns@gmail.com

William Takahiro
Maruyama
University of Sao Paulo
Av. Arlindo Bettio, 1000
Sao Paulo, SP, Brazil
takahiro.mw@gmail.com

Jamison José da Silva
Lima
University of Sao Paulo
Av. Arlindo Bettio, 1000
Sao Paulo, SP, Brazil
jamison.lima@gmail.com

RESUMO

Predizer novos relacionamentos dentro de um grupo social é uma tarefa complexa, porém extremamente útil para potencializar ou maximizar colaborações por meio da indicação de quais seriam as parcerias mais promissoras. Nas redes sociais acadêmicas, a predição de relacionamentos é tipicamente utilizada para tentar identificar potenciais parceiros no desenvolvimento de um projeto e/ou coautores para a publicação de um artigo. Este artigo apresenta um sistema que combina técnicas de inteligência artificial com o estado da arte das métricas de predição de relacionamentos em redes sociais. O sistema resultante foi testado usando dados reais de pesquisadores em Ciência da Computação e atingiu uma precisão superior a 99,5% na predição de novas coautorias.

Palavras-Chave

predição de relacionamentos, redes sociais, redes acadêmicas

ABSTRACT

The prediction of new relationships in a social network is a complex and extremely useful task to enhance or maximize collaborations by indicating what the most promising partnerships are. In academic social networks, prediction of relationships is typically used to try to identify potential partners in the development of a project and/or co-authors for publishing papers. This paper presents a system that combines artificial intelligence techniques with the state-of-the-art metrics for link prediction. The resulting system was

tested using real data from Computer Science researchers and achieved a precision above 99.5% in the coauthorship prediction.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Information retrieval

General Terms

Information Retrieval

Keywords

link prediction, social networks, academic networks

1. INTRODUÇÃO

A predição de novos relacionamentos dentro de uma rede social é uma tarefa que ganhou bastante destaque nos últimos anos [25, 21, 30, 20, 27, 6], pois serve para desde encontrar amigos que ainda não estavam ligados em uma rede social online [35, 34, 31, 15, 36, 32], até para potencializar a realização de trabalhos em empresas ou na academia [23, 14, 5].

Alguns dos fatores que tornam esta tarefa complexa são: a identificação de quais atributos individuais (relacionados, por exemplo, ao perfil ou currículo das pessoas) serão utilizados; especificação ou seleção de métricas estruturais de redes sociais a serem usadas; utilização de estratégias para combinar estes atributos e métricas de forma a possibilitar a predição; e uso correto do fator temporal para maximizar a precisão e a sensibilidade¹ da solução produzida.

Nas redes sociais acadêmicas, a predição de relacionamentos tem sido utilizada principalmente para a predição de coautorias [13, 7, 16, 18, 26, 28], atividade que indica se um

¹A *sensibilidade* ou *revocação* corresponde a quantidade dos elementos corretamente classificados de forma positiva (verdadeiro-positivos) dividida pela quantidade total de elementos pertencentes a classe positiva.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

par de pesquisadores poderá/deverá colaborar na produção de um artigo. Este tipo de predição pode, assim, otimizar a produção destes pesquisadores por meio da indicação de pesquisadores cujas parcerias são mais promissoras.

O problema de predição de relacionamentos, por sua vez, pode ser dividido em predição de relacionamentos novos/inéditos (isto é, prever quais pares de pessoas que nunca se relacionaram numa rede social irão começar a se relacionar) ou no problema geral de predição de relacionamentos (predizer que pares de pessoas irão se relacionar independentemente delas já terem ou não se relacionado). Tipicamente, a expressão “predição de relacionamentos” (ou de *links*) se refere ao primeiro problema (predição de relacionamentos novos/inéditos).

Este artigo utiliza um sistema gerenciador de workflows científicos (*Scientific Workflow Management System (SWMS)*) para a construção de um sistema geral e modular de predição de relacionamentos em redes sociais. A entrada básica para o sistema é um arquivo contendo a lista de relacionamentos criados ao longo do tempo entre um conjunto de pessoas. Adicionalmente, foram desenvolvidos módulos específicos para a extração de informações de currículos Lattes de forma a se extrair informações da rede social acadêmica de um conjunto de pessoas e, com base nessas informações, realizar a predição de coautorias.

O restante deste artigo está organizado da seguinte maneira. A Seção 2 apresenta os trabalhos correlatos. A Seção 3 contém a descrição da metodologia utilizada. Na Seção 4 é apresentado um estudo de caso sobre predição de coautorias em redes sociais acadêmicas. Por fim, a Seção 5 contém as conclusões e os trabalhos futuros.

2. TRABALHOS CORRELATOS

Nos últimos anos diversos trabalhos foram publicados sobre a predição de relacionamentos. De um modo geral, estes trabalhos podem ser divididos em três grupos: aqueles que utilizam apenas características da rede social (ou mais especificamente, do grafo que representa a rede social) [13, 3, 14, 16]; aqueles que propõem a utilização de atributos (primitivos ou derivados) específicos do domínio no qual a predição irá ocorrer [35]; e sistemas híbridos que combinam estes dois aspectos [24, 7, 34, 18].

Ao se analisar os trabalhos que propõem a utilização de um ou mais atributos (ou mesmo a criação de atributos derivados) é possível observar uma gama muito grande e diversificada de atributos. Mesmo ao se restringir o domínio para predição de coautorias o número de correlatos ainda é elevado [13, 7, 16, 18, 26, 28, 5, 4].

A partir de uma revisão sistemática realizada sobre este tema, foram selecionados os atributos e as métricas que obtiveram os melhores desempenhos para serem aplicadas no problema de predição de novos relacionamentos. Estas métricas serão apresentadas na próxima seção.

Além disso, foi encontrado um trabalho específico sobre predição de coautorias usando dados de currículos Lattes [7]. Nesse trabalho, todos os atributos utilizados são oriundos de informações dos currículos da Plataforma Lattes. Esse trabalho analisou a predição de coautorias dos docentes dos programas brasileiros de pós-graduação em Ciência da Computação e utilizou os seguintes atributos (para cada par de docentes): artigos publicados em conjunto (em periódicos e em conferências); relação de orientação entre os docentes; existência de orientadores em comum; existência de orien-

tandos em comum; se os dois pertencem ao mesmo programa de pós-graduação; e quantos vizinhos em comum eles possuem na rede de coautoria. Além disto, estes atributos foram divididos em períodos de tempo: os dados de 1971 a 2000 foram considerados passados; de 2001 a 2005, dados atuais e o objetivo dos autores era predizer as coautorias que ocorreram de 2006 a 2010. O problema de predição de relacionamentos foi abordado como um problema de classificação no qual, dado o vetor de atributos correspondente às informações de um par de docentes, diferentes algoritmos de inteligência artificial foram usados para classificar estes docentes em “serão coautores” ou “não serão coautores”. Esta estratégia permitiu aos autores uma acurácia de 99,709% na predição, valor bastante expressivo, porém é importante observar, como destacado pelos autores, que dado um par arbitrário de docentes da amostra utilizada, há 99,6% de chance deles não serem coautores (um sistema que classificasse todos os pares como “não serão coautores” acertaria em 99,6% dos casos). O sistema não utilizou nenhuma estratégia para balanceamento de dados e, provavelmente por isto, a principal limitação da abordagem utilizada nesse trabalho foi sua incapacidade de predizer novas coautorias, isto é, o sistema se mostrou muito bom em identificar quais pares de docentes iriam publicar novamente juntos, mas ao tratar especificamente do problema de identificar quais serão os novos coautores a acurácia ficou abaixo daquela que seria obtida por um classificador que sempre classificasse os pares como “não serão coautores”.

No presente artigo será apresentado um sistema de recomendação de relacionamentos em redes sociais utilizando uma abordagem robusta, por meio da combinação de diferentes atributos estruturais de rede (oriundos dos trabalhos correlatos analisados) e, para a realização do estudo de caso em redes sociais acadêmicas, também serão utilizados diversos atributos específicos desse domínio. Os diferentes atributos serão combinados utilizando-se técnicas de inteligência artificial a fim de se predizer tanto relacionamentos inéditos como a reincidência de relacionamentos. Tanto os atributos a serem utilizados como o algoritmo de inteligência artificial que irá combiná-los podem ser escolhidos pelo usuário. Adicionalmente, há um parâmetro específico para indicar se o usuário pretende prever apenas novos relacionamentos ou também prever a reincidência de relacionamentos (por exemplo, se um par de docentes que já colaborou na produção de um artigo irá colaborar novamente).

Neste trabalho a predição de relacionamentos também será abordada como um problema de classificação e para cada par de pessoas será atribuída uma das seguintes classes “possuirão relacionamento” ou “não possuirão relacionamento”.

O sistema desenvolvido neste trabalho opera sobre um SWMS, isto permite que novos módulos sejam facilmente incorporados ao sistema além de possibilitar a execução paralela ou distribuída das diferentes atividades/módulos do workflow. Há diversos SWMSs disponíveis na literatura [1, 33, 2, 17]. Optou por utilizar uma das extensões do SWMS chamado WOODSS [29] pelos seguintes motivos: (i) o sistema foi desenvolvido no Brasil e possui seu código fonte disponível na linguagem de programação Java; (ii) o sistema permite que as atividades básicas (módulos) dos workflows sejam de diferentes tipos (por exemplo, serviços Web ou métodos na linguagem Java) [9]; (iii) o sistema já foi estendido para diferentes aplicações [11, 8]; e (iv) a versão utilizada do

SWMS possui um módulo de execução distribuída que pode utilizar computação voluntária (quando disponível), o que possibilita a execução eficiente dos workflows [12].

3. METODOLOGIA

O trabalho apresentado neste artigo foi organizado em quatro atividades: revisão da literatura correlata (sumarizada na Seção 2); especificação do sistema; desenvolvimento dos módulos; e realização de um estudo de caso.

3.1 Especificação do sistema

Conforme apresentado, o sistema de predição foi construído sobre um SWMS. Desta forma, o sistema se beneficiou de diversas características desse SWMS (tais como, execução paralela e distribuída, interface gráfica para construção, edição dos workflows e preenchimento de parâmetros das atividades). Desta forma, o sistema de predição precisará, minimamente, conter os módulos de processamento (atividades de workflow) necessários para a realização da predição. A Figura 3.1 apresenta a organização do sistema em seus módulos. Detalhes sobre a versão atual do SWMS bem como exemplos de workflows são apresentados na dissertação de mestrado de Caio Nascimento Santiago [12].



Figura 1: Módulos do Sistema Implementado

O sistema é composto de seis módulos, que serão definidos a seguir e terão algumas características detalhadas na próxima subseção. O módulo *Extrator de Atributos Estruturais* recebe como entrada um conjunto de triplas que indicam os relacionamentos existentes ao longo dos anos, no formato $\langle id1, id2, ano \rangle$ (identificador da pessoa 1, da pessoa 2 e ano em que ocorreu o relacionamento entre estas pessoas) e extrai um conjunto de métricas/atributos estruturais de rede que podem ser utilizadas para a predição de relacionamentos em diferentes domínios (por exemplo, redes sociais de amizade, de trabalho, etc).

O módulo *Extrator de Atributos Específicos* foi concebido para processar dados de currículos Lattes. Ele recebe como parâmetro uma lista de identificadores de currículos e extrai diferentes características específicas a serem utilizadas na predição. Além disso, quando este módulo é utilizado ele produz as triplas utilizadas como entrada pelo módulo *Extrator de Atributos Estruturais* (que neste caso corresponderão às coautorias realizadas entre os pesquisadores possuidores dos currículos que estão sendo analisados). A maioria dos módulos recebe ainda três intervalos de tempo como parâmetros de entrada, que são chamados de *passado*, *presente* e *futuro*. O *futuro* indica o período de tempo que inicialmente será utilizado para a validação do treinamento do sistema de predição. Corresponde a um intervalo de tempo cujos dados de entrada são conhecidos, mas serão utilizados pelo *Preditor* durante o treinamento e validação dos resultados. O *presente* constitui o conjunto de dados mais recente exceto os dados do *futuro*. Os dados do *passado* são opcionais e visam a agregar informação que pode ser útil para a predição mas que não é muito recente. Esta separação entre *passado*

e *presente* ou a utilização de ponderações diferentes entre dados mais ou menos recentes tende a melhorar a acurácia na predição de relacionamentos [34, 3].

O módulo *Concatenador de Atributos* é responsável por concatenar os diferentes atributos de entrada de forma a produzir um conjunto de dados padronizado a ser utilizado pelos próximos módulos. Este módulo recebe um conjunto de tuplas no formato $\langle id1, id2, atributo, valor \rangle$ correspondendo ao valor de um dado atributo para o par de pessoas identificadas por *id1* e *id2* e gera uma matriz na qual cada linha corresponde a cada par de pessoas da amostra utilizada e cada coluna corresponde a um atributo.

O módulo *Filtro Horizontal de Dados* recebe a matriz de pares de pessoas (linhas) por atributos (colunas) e elimina as linhas que não tenham nenhum valor diferente de nulo (ou de acordo com outras condições detalhadas na próxima subseção). As linhas filtradas/excluídas tipicamente são rotuladas como “não possuirão relacionamento” e são excluídas por não agregarem informação ao *Preditor* que irá tentar prever os relacionamentos para os demais pares de pessoas (isto é, as linhas restantes da matriz). Há uma função adicional deste módulo: preparar a matriz de dados para o problema de predição de (novos) relacionamentos ou para o problema que envolve também a predição da reincidência de relacionamentos. Um dos parâmetros passado pelo usuário determinará o problema que se deseja abordar e a diferença na execução deste módulo é a seguinte: para o caso de predição de (novos) relacionamentos o módulo também deverá excluir as linhas que correspondem a pares de pessoas que já se relacionam no *presente*.

O módulo *Balanceador da Amostra* recebe a matriz resultante da execução do *Filtro Horizontal de Dados* e produz como saída uma matriz com dados balanceados segundo o atributo classe (isto é, “possuirão relacionamento” ou “não possuirão relacionamento”). Este módulo é particularmente importante para o problema de predição de relacionamentos, pois em redes sociais a existência ou não de relacionamentos tipicamente produz um conjunto de dados bastante desbalanceado. Isto é, dado um par arbitrário de pessoas é muito mais provável que estas pessoas não irão se relacionar do que o contrário. Esse tipo de desbalanceamento, se não tratado, pode afetar significativamente o desempenho do *Preditor* (que neste trabalho é um classificador), afetando principalmente a sensibilidade ou revocação da solução [22].

O módulo *Preditor* recebe a matriz resultante da execução do *Balanceador da Amostra* e produz como saída a classificação dos pares de pessoas indicando se possuirão ou não um relacionamento. Este módulo utiliza os dados do *passado* e do *presente* para treinamento e teste (verificando a precisão em relação ao *futuro*) e, em seguida, realiza a predição dos dados mais recentes (do *futuro*) utilizando os dados do *presente* como dados anteriores (isto é, como do *passado*).

Características importantes da implementação atual dos módulos serão descritas na próxima subseção.

3.2 Desenvolvimento dos módulos

Todos os módulos desenvolvidos neste trabalho foram implementados na linguagem de programação Java, o que permitiu sua automática incorporação como atividades de workflow no SWMS utilizado. A seguir são detalhadas algumas características de cada módulo e em especial as técnicas ou algoritmos utilizados.

A versão atual do módulo *Extrator de Atributos Estru-*

Tabela 1: Atributos Estruturais Utilizados

Atributo	Descrição
classe	Atributo que assume o valor “possuirão relacionamento” ou “não possuirão relacionamento” com base nos dados do intervalo de tempo chamado neste trabalho de <i>futuro</i> .
CN	Common Neighbors - número de vizinhos em comum no grafo correspondente à rede social.
SAL	Salton Index - índice que mede a coocorrência de dois elementos dividida pela raiz quadrada da multiplicação da ocorrência de cada elemento. Em redes sociais pode ser usado para medir relação entre o número de vizinhos que duas pessoas têm em comum dividido pela raiz quadrada da multiplicação do número de vizinhos de cada um.
JAC	Jaccard's coefficient - índice que mede a similaridade entre dois conjuntos dividindo o número de elementos da intersecção dos dois conjuntos pelo número de elementos da união (por exemplo, número de vizinhos em comum dividido pela união dos vizinhos de duas pessoas).
AA	Adamic-Adar - índice que atribui peso na relação de duas pessoas favorecendo as relações entre pessoas que possuem poucos relacionamentos (o peso do relacionamento é calculado pela somatória de 1 dividido pelo logaritmo do número de relacionamentos [grau] dos vizinhos em comum destas duas pessoas).
RA	Resource Allocation - índice que atribui peso na relação de duas pessoas favorecendo as relações entre pessoas que possuem poucos relacionamentos (o peso do relacionamento é calculado pela somatória de 1 dividido pelo número de relacionamentos [grau] dos vizinhos em comum destas duas pessoas).
SOR	Sørensen Index - índice calculado como sendo duas vezes a intersecção entre dois conjuntos dividida pela soma dos elementos de cada conjunto (por exemplo, número de vizinhos em comum dividido pelo número de vizinhos da primeira pessoa mais o número de vizinhos da segunda).
HPI	Hub Promoted Index - índice calculado pela divisão do número de elementos da intersecção de dois conjuntos dividido pelo número mínimo de elementos entre estes dois conjuntos (por exemplo, número de vizinhos em comum de duas pessoas dividido pelo número mínimo de vizinhos destas pessoas).
HDI	Hub Depressed Index - índice calculado pela divisão do número de elementos da intersecção de dois conjuntos dividido pelo número máximo de elementos entre estes dois conjuntos (por exemplo, número de vizinhos em comum de duas pessoas dividido pelo número máximo de vizinhos destas pessoas).
LHM	Leicht-Holme-Newman Index - índice calculado pelo número de elementos da intersecção de dois conjuntos dividido pelo produto do número de elementos de cada conjunto (por exemplo, número de vizinhos em comum dividido pela multiplicação do número de vizinhos de duas pessoas).
PA	Preferential Attachment - índice dado pelo produto entre o número de elementos de dois conjuntos (por exemplo, produto do número de vizinhos de duas pessoas).
KATZ0,05 KATZ0,005 KATZ0,0005	Katz é um índice calculado de maneira iterativa para estimar a influência de um par de nós em uma rede considerando-se os caminhos existentes entre os nós. Para este cálculo existe a necessidade da definição de uma constante Beta. Os valores utilizados foram: 0,05 ; 0,005 ; e 0,0005.
SP	Shortest Path - caminho mínimo entre dois nós da rede.

turais possui 14 atributos (ou características) estruturais calculados (além do atributo *classe*). Estes atributos são oriundos de métricas estabelecidas por trabalhos correlatos que obtiveram bons desempenhos na predição de relacionamentos [25, 21, 30, 20, 27, 6]. Todos estes atributos são calculados utilizando-se apenas o grafo que representa a rede de relacionamentos. A Tabela 1 apresenta o nome e a descrição destes atributos. Observa-se que foram utilizados três atributos derivados do índice Katz. A maioria desses atributos faz sentido apenas para pares de pessoas pertencentes ao mesmo componente conexo da rede, para os demais pares alguns atributos possuem valores padrões ou simplesmente nenhum valor é produzido para o respectivo atributo.

O módulo *Extrator de Atributos Específicos* recebe como entrada a lista de identificadores de currículos Lattes, copia da internet a versão HTML dos respectivos currículos e produz 19 atributos, conforme pode ser visto na Tabela 2. Este módulo só é utilizado para a predição de coautorias que utilizem dados de currículos Lattes. Para o problema geral de predição de relacionamentos utiliza-se apenas o módulo *Extrator de Atributos Estruturais*. Os dois primeiros atributos são relacionados à classe. O primeiro, *coautorias a serem preditas* contém o número de coautorias que ocorrerão no período *futuro* (ou seja, aquelas que deseja-se prever). O segundo contém a *classe* propriamente dita (no caso, uma

indicação se o respectivo par de pessoas colaborará ou não em uma coautoria).

Todos os atributos específicos, exceto os seis últimos, foram extraídos do trabalho de Digiampietri *et al.*, 2013 [7], sendo que, para se calcular o atributo, *programas em comum* este módulo precisa receber uma tabela que relacione cada pesquisador ao programa de pós-graduação (ou departamento) em que atua, pois esta informação nem sempre é explícita nos currículos Lattes. Dos atributos específicos incluídos no presente trabalho, quatro se referem à produção individual de cada uma das pessoas envolvidas no par de pessoas em análise. Já os dois últimos são *distância* (distância euclidiana entre as coordenadas do endereço profissional de cada par de pessoas) e *subáreas em comum* (correspondendo a contagem de subáreas em comum declaradas pelas duas pessoas no campo “Áreas de atuação” de seus currículos Lattes). Destaca-se ainda o atributo *vizinhos em comum*, apesar de ser semelhante ao atributo estrutural *CN* a diferença é que o primeiro contém todos os “vizinhos” na rede social acadêmica formada pelos diferentes tipos de relacionamento (coautoria, participação em projetos, orientação, etc), já o segundo contém a contagem de vizinhos apenas para a rede (ou grafo) de um relacionamento específico (por exemplo, o de coautoria).

O módulo *Concatenador de Atributos* apenas realiza a

Tabela 2: Atributos Específicos Utilizados

Atributo	Descrição
coautorias a serem preditas	Quantidade de artigos completos publicados em coautoria pelo par de pesquisadores em análise em conferências ou em periódicos no período <i>futuro</i> .
classe	Atributo que assume o valor “possuirão relacionamento” caso o atributo “coautorias a serem preditas” seja maior que zero e, caso contrário, “não possuirão relacionamento”.
periódicos anterior	Quantidade de artigos publicados em periódicos em coautoria pelo par de pesquisadores no período <i>passado</i> .
conferências anterior	Quantidade de artigos completos publicados em conferências em coautoria pelo par de pesquisadores no período <i>passado</i> .
periódicos atual	Quantidade de artigos publicados em periódicos em coautoria pelo par de pesquisadores no período <i>presente</i> .
conferências atual	Quantidade de artigos completos publicados em conferências em coautoria pelo par de pesquisadores no período <i>presente</i> .
orientação anterior	Atributo que recebe o valor 1 (um) caso um dos pesquisadores tenha sido orientador do outro no período <i>passado</i> , ou 0 (zero) caso contrário.
orientação atual	Atributo que recebe o valor 1 (um) caso um dos pesquisadores tenha sido orientador do outro no período <i>presente</i> , ou 0 (zero) caso contrário.
orientação em andamento	Atributo que recebe o valor 1 (um) caso um dos pesquisadores seja orientador, em uma orientação em andamento, no período <i>presente</i> , ou 0 (zero) caso contrário.
orientadores em comum	Quantidade de orientadores e coorientadores que foram orientadores dos dois pesquisadores em análise.
orientandos em comum	Quantidade de orientandos e coorientandos que foram orientados pelos dois pesquisadores em análise.
vizinhos em comum	Quantidade de vizinhos em comum entre os dois pesquisadores na rede social acadêmica (incluindo os diferentes relacionamentos: coautoria, orientação, etc).
programas em comum	Atributo que recebe o valor 1 (um) caso os dois pesquisadores pertençam ao mesmo programa de pós-graduação, ou 0 (zero) caso contrário.
artigos periódico1	Quantidade de artigos publicados em periódicos no período <i>presente</i> pela pessoa 1.
artigos anais1	Quantidade de artigos completos publicados em anais de conferências no período <i>presente</i> pela pessoa 1.
artigos periódico2	Quantidade de artigos publicados em periódicos no período <i>presente</i> pela pessoa 2.
artigos anais2	Quantidade de artigos completos publicados em anais de conferências no período <i>presente</i> pela pessoa 2.
distância	Distância geográfica entre os endereços profissionais dos dois pesquisadores.
subáreas em comum	Número de subáreas de atuação que os dois pesquisadores possuem em comum.

concatenação dos atributos recebidos como entrada, executando ou não a normalização dos valores de cada atributo (de acordo com um parâmetro de entrada) e produz uma matriz na qual cada linha corresponde a um par de pessoas e cada coluna corresponde a um atributo.

O módulo *Filtro Horizontal de Dados* possui três formas de filtragem de dados atualmente implementadas. Na primeira são descartadas as linhas (correspondendo a um par de pessoas) nas quais todos os valores de atributos são nulos. Na segunda são eliminadas todas as linhas cujos pares de pessoas não estão no mesmo componente conexo. Na terceira forma o usuário passa como parâmetro uma expressão booleana que será aplicada sobre cada uma das linhas e indicará se a linha deverá ou não ser descartada.

A versão atual do módulo *Balanceador da Amostra* permite apenas uma estratégia de balanceamento, o *Random Oversampling*. Esta estratégia adiciona aleatoriamente elementos da classe minoritária até que o número de elementos dessa classe se iguale ao número de elementos da classe majoritária.

Conforme apresentado, o problema de predição é, neste trabalho, visto como um problema de classificação em in-

teligência artificial. O módulo *Preditor*, além de receber o conjunto de dados a ser utilizado no treinamento e teste também receberá um parâmetro indicando qual o planejador a ser utilizado, o caminho (dentro do sistema de arquivos) onde esse planejador se encontra e a estratégia de validação a ser utilizada. Ao invés de se implementar um ou mais classificadores, optou-se por utilizar os classificadores disponíveis no ambiente Weka [19]. Com base nos parâmetros de entrada, o *Preditor* utiliza de reflexão (*reflection*²) para invocar o classificador e obter o resultado da classificação/predição.

4. ESTUDO DE CASO

O estudo de caso realizado neste trabalho envolveu a predição de coautorias. A amostra selecionada corresponde aos 657 docentes permanentes dos programas de pós-graduação em Ciência da Computação com doutorado e/ou mestrado acadêmico que atuaram nos triênios 2004-2006 e 2007-2009. As três principais motivações para o uso deste conjunto de dados são: (i) disponibilidade dos dados; (ii) validação: estes dados já foram utilizados para a predição de coautorias

²<http://docs.oracle.com/javase/tutorial/reflect/>

Tabela 3: Resultados da Predição de Novas Coautorias

	Taxa de verdadeiro-positivos	Taxa de falso-positivos	Precisão	Revocação	F-Measure	Área ROC
Não serão coautores	0,992	0	1	0,992	0,996	1
Serão coautores	1	0,008	0,992	1	0,996	1
Média ponderada	0,996	0,004	0,996	0,996	0,996	1

em [7], então será possível comparar os resultados da solução proposta com os resultados obtidos pelos autores; (iii) ser um amostra de interesse: pelo fato da amostra conter dados de docentes permanentes dos programas de pós-graduação de uma única área do conhecimento há diversos tipos de relacionamentos pertinentes para a predição de coautorias: relações de orientação; de trabalho (num mesmo programa de pós-graduação); e de coautorias [10].

Os parâmetros utilizados para a realização do estudo de caso são: intervalo *passado* envolveu os anos de 1971 a 2000; o *presente* de 2001 a 2005 e o *futuro* de 2006 a 2010. Na filtragem horizontal dos dados foram excluídos os pares de elementos para os quais mais de metade dos atributos possuem valores nulos. Adicionalmente foi passado o parâmetro que indica que o problema a ser tratado é o de predição de relacionamentos novos/inéditos. O classificador utilizado foi o *Rotation Forest* que experimentalmente havia apresentado os melhores resultados para diferentes conjuntos de dados e a estratégia de validação selecionada foi *10-fold-cross-validation*.

Ao se combinar os 657 pesquisadores da amostra dois a dois obtêm-se 215.496 pares. Nesta amostra, o número de pares que efetivamente serão coautores (no *futuro*) é muito menor (apenas 804 pares). Dos 215.496 pares apenas 10.996 foram mantidos após a execução do filtro horizontal de dados. Os demais 204.500 pares (excluídos pelo filtro) são automaticamente classificados como “não possuirão relacionamento” (que neste estudo de caso significa que não serão coautores no *futuro*). Obviamente esta classificação poderia implicar em alguns falso-negativos, porém nenhum destes 204.500 pares foram coautores entre 2006 e 2010. Dos 10.996 pares de docentes restantes, 10.094 pares não haviam colaborado em nenhuma publicação antes de 2005 (ou seja, são candidatos a predição de coautorias novas/inéditas).

A Tabela 3 apresenta os resultados da execução do classificador *Rotation Forest* utilizando a estratégia *10-fold-cross-validation* sobre o conjunto de dados balanceado. As métricas utilizadas para a apresentação dos resultados são: *Taxa de verdadeiro-positivos*, isto é, a quantidade de elementos corretamente classificados como pertencentes à classe dividida pela quantidade de elementos que são da classe; *Taxa de falso-positivos*, isto é, a quantidade de elementos erroneamente classificados como pertencentes à classe dividida pela quantidade de elementos pertencentes à outra classe; *Precisão*, isto é, o número de elementos verdadeiro-positivos dividido pela soma dos verdadeiro-positivos com os falso-positivos; *Revocação*, isto é, a quantidade de elementos classificados como pertencente à classe dividido pela quantidade total de elementos da classe; *F-Measure*, que corresponde a duas vezes o valor da precisão vezes a revocação dividido pela soma da precisão com a revocação; e *Área ROC*.

Conforme pode ser observado na Tabela 3 os resultados foram bastante satisfatórios atingindo uma precisão média

acima de 99% para o conjunto balanceado e, ao se considerar também os elementos excluídos pelo filtro e que são classificados como “não serão coautores” a precisão sobe para 99,96% (considerando todo o conjunto de dados). De fato, ao se comparar o resultado da predição para o período *futuro* com o que efetivamente ocorreu (isto é, se os pares de pessoas realmente colaboraram em uma publicação nesse período) a predição errou apenas em 81 dos 215.496 pares de pessoas e o erro foi dizer que estes 81 pares seriam coautores quando na verdade não foram. Pode-se atribuir os resultados obtidos a três características principais. A primeira é que foram utilizados diferentes atributos estruturais encontrados na literatura com bons desempenhos para a predição de relacionamentos. Adicionalmente, foram utilizados atributos específicos do domínio (alguns desenvolvidos especificamente neste trabalho e outros encontrados na literatura). Por fim, o conjunto de dados utilizados nos testes é composto por elementos relativamente homogêneos (todas as pessoas são professores orientadores permanentes por ao menos dois triênios em programas de pós-graduação em Ciência da Computação) o que deve ter simplificado o processo de classificação.

Para uma comparação mais detalhada dos resultados, são apresentados na Tabela 4 os resultados médios de alguns atributos individuais, do conjunto de atributos específicos utilizados em trabalho correlato [7] e, na última linha, a média dos resultados da solução proposta. Todos estes resultados calculados para o conjunto balanceado de dados, utilizando o classificador *Rotation Forest* e a estratégia de validação *10-fold-cross-validation*. Observa-se nesta tabela que os resultados de atributos individuais estão bem aquém dos resultados utilizando-se todos os atributos (última linha) ou ao menos um conjunto de atributos específicos (penúltima linha). Este fato confirma a premissa deste trabalho de que um sistema de predição que combine diferentes atributos/características tende a ser mais preciso e o robusto.

A Tabela 5 apresenta os resultados, utilizando-se a mesma amostra, para o problema que considera tanto a predição de novas coautorias como a predição da reincidência de coautorias. Os resultados utilizaram dados balanceados, o classificador *Rotation Forest* e a estratégia de validação *10-fold-cross-validation*. A precisão, se calculada para todo conjunto de entrada (incluindo os elementos que foram excluídos pelo filtro) é de 99,90%.

5. CONCLUSÕES

Este trabalho apresentou um sistema para a predição de relacionamentos em redes sociais baseado em estratégias de classificação em inteligência artificial para combinar atributos que podem ser estruturais da rede, mas também específicos do domínio.

Adicionalmente foi realizado um estudo de caso real utilizando-se dados da Plataforma Lattes de professores de pro-

Tabela 4: Comparações de Resultados

	Taxa de verdadeiro-positivos	Taxa de falso-positivos	Precisão	Revocação	F-Measure	Área ROC
Subáreas em comum	0,547	0,453	0,547	0,547	0,547	0,542
Vizinhos em comum	0,524	0,476	0,56	0,524	0,438	0,522
PA	0,598	0,402	0,599	0,598	0,597	0,611
Distância	0,596	0,404	0,636	0,596	0,565	0,645
CN	0,61	0,39	0,639	0,61	0,588	0,606
Katz 0,05	0,647	0,353	0,685	0,647	0,628	0,707
Específicos[7]	0,993	0,007	0,993	0,993	0,993	1
Solução Proposta	0,996	0,004	0,996	0,996	0,996	1

Tabela 5: Resultados da Predição de Coautorias (Novas e Reincidentes)

	Taxa de verdadeiro-positivos	Taxa de falso-positivos	Precisão	Revocação	F-Measure	Área ROC
Não serão coautores	0,98	0	1	0,98	0,99	0,999
Serão coautores	1	0,02	0,98	1	0,99	0,999
Média ponderada	0,99	0,01	0,99	0,99	0,99	0,999

gramas de pós-graduação em ciência da computação, no qual foi obtida uma precisão para o problema de predição de (novas/inéditas) coautorias acima de 99,5% para os dados filtrados e de 99,96% considerando-se o conjunto total de dados.

Também foi possível verificar que a abordagem proposta pode ser utilizada para o problema geral de predição de coautorias (e não apenas de coautorias inéditas), atingindo a precisão de 99% para o conjunto filtrado de dados e 99,9% para o conjunto total de dados.

Como trabalhos futuros pretende-se desenvolver uma interface gráfica específica para o sistema de predição que atualmente utiliza a interface genérica de um sistema de gerenciamento de workflows científicos. Adicionalmente, pretende-se testar e incorporar outras características para a predição de relacionamentos.

Agradecimentos

O trabalho desenvolvido neste artigo foi parcialmente financiado pela CAPES, CNPq e FAPESP.

6. REFERÊNCIAS

- [1] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock. Kepler: An extensible system for design and execution of scientific workflows. In *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*, pages 423–424, Washington, DC, USA, 2004.
- [2] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Managing the evolution of dataflows with VisTrails. In *Proceedings of the 22nd International Conference on Data Engineering Workshops*, page 71, 2006.
- [3] W. Cukierski, B. Hamner, and B. Yang. Graph-based features for supervised link prediction. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1237–1244, July 2011.
- [4] P. da Silva Soares and R. Bastos Cavalcante Prudencio. Time series based link prediction. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–7, June 2012.
- [5] H. de Sa and R. Prudencio. Supervised link prediction in weighted networks. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 2281–2288, July 2011.
- [6] Y. Dhote, N. Mishra, and S. Sharma. Survey and analysis of temporal link prediction in online social networks. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pages 1178–1183, Aug 2013.
- [7] L. Digiampietri, C. Santiago, and C. Alves. Predição de coautorias em redes sociais acadêmicas: um estudo exploratório em ciência da computação. In *CSBC-BraSNAM 2013*, jul 2013.
- [8] L. Digiampietri, B. Teodoro, C. Santiago, G. Oliveira, and J. Araújo. Um sistema de informação extensível para o reconhecimento automático de libras. In *SBSI 2012 - Trilhas Técnicas (Technical Tracks)*, may 2012.
- [9] L. A. Digiampietri, J. C. Araujo, E. H. Ostroski, C. R. N. Santiago, and J. J. Perez-Alcazar. Combinando workflows e semântica para facilitar o reuso de software. *Revista de Informática Teórica e Aplicada: RITA*, 20:73–89, 2013.
- [10] L. A. Digiampietri, J. P. Mena-Chalco, P. O. S. Vaz de Melo, A. P. R. Malheiro, D. N. O. Meira, L. F. Franco, and L. B. Oliveira. Brax-ray: An x-ray of the brazilian computer science graduate programs. *PLoS ONE*, 9(4):e94541, 04 2014.
- [11] L. A. Digiampietri, V. M. Pereira, C. I. Costa, G. J. dos Santos Junior, F. M. Stefanini, and C. R. Santiago. An extensible framework for genomic and metagenomic analysis. In S. Campos, editor, *Advances in Bioinformatics and Computational Biology*, volume 8826 of *Lecture Notes in Computer Science*, pages 1–8. Springer International Publishing, 2014.
- [12] C. R. do Nascimento Santiago. Desenvolvimento de

- um ambiente de computação voluntária baseado em computação ponto-a-ponto. Master's thesis, Universidade de Sao Paulo, 2015.
- [13] Y. Dong, Q. Ke, J. Rao, and B. Wu. Predicting missing links via local feature of common neighbors. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, volume 2, pages 1038–1042, July 2011.
- [14] Y. Dong, J. Tang, S. Wu, J. Tian, N. Chawla, J. Rao, and H. Cao. Link prediction and recommendation across heterogeneous social networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 181–190, Dec 2012.
- [15] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*, pages 73–80, Oct 2011.
- [16] S. Gao, L. Denoyer, and P. Gallinari. Link prediction via latent factor blockmodel. In *Proceedings of the 21st International Conference Companion on World Wide Web, WWW '12 Companion*, pages 507–508, New York, NY, USA, 2012. ACM.
- [17] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.
- [18] J. Guo and H. Guo. Multi-features link prediction based on matrix. In *Computer Design and Applications (ICCD), 2010 International Conference on*, volume 1, pages V1–357–V1–361, June 2010.
- [19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.
- [20] M. Hasan and M. Zaki. A survey of link prediction in social networks. In C. C. Aggarwal, editor, *Social Network Data Analytics*, pages 243–275. Springer US, 2011.
- [21] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.
- [22] H. He, Y. Bai, E. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1322–1328, June 2008.
- [23] C.-J. Hsieh, M. Tiwari, D. Agarwal, X. L. Huang, and S. Shah. Organizational overlap on social networks and its applications. In *Proceedings of the 22Nd International Conference on World Wide Web, WWW '13*, pages 571–582, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- [24] J. Kunegis, J. Preusse, and F. Schwagereit. What is the added value of negative links in online social networks? In *Proceedings of the 22Nd International Conference on World Wide Web, WWW '13*, pages 727–736, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- [25] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM '03*, pages 556–559, New York, NY, USA, 2003. ACM.
- [26] Z. Lin, X. Yun, and Y. Zhu. Link prediction using benefitranks in weighted networks. In *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01, WI-IAT '12*, pages 423–430, Washington, DC, USA, 2012. IEEE Computer Society.
- [27] L. Lu and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150 – 1170, 2011.
- [28] M. Makrehchi. Social link recommendation by learning hidden topics. In *Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11*, pages 189–196, New York, NY, USA, 2011. ACM.
- [29] C. Medeiros, J. Perez-Alcazar, L. Digiampietri, G. Pastorello, A. Santanche, R. Torres, E. Madeira, and E. Bacarin. WOODSS and the Web: Annotating and Reusing Scientific Workflows. *ACM SIGMOD Record*, 34(3):18–23, 2005.
- [30] T. Murata and S. Moriyasu. Link prediction based on structural properties of online social networks. *New Generation Computing*, 26(3):245–257, 2008.
- [31] C. Perez, B. Birregah, and M. Lemerrier. The multi-layer imbrication for data leakage prevention from mobile devices. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, pages 813–819, June 2012.
- [32] D. Quercia and L. Capra. Friendsensing: Recommending friends using mobile phones. In *Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09*, pages 273–276, New York, NY, USA, 2009. ACM.
- [33] I. Taylor, M. Shields, I. Wang, and A. Harrison. Visual grid workflow in triana. *Journal of Grid Computing*, 3(3-4):153–169, 2005.
- [34] Y. Tian, Q. He, Q. Zhao, X. Liu, and W.-c. Lee. Boosting social network connectivity with link revival. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, pages 589–598, New York, NY, USA, 2010. ACM.
- [35] V. Vasuki, N. Natarajan, Z. Lu, and I. S. Dhillon. Affiliation recommendation using auxiliary networks. In *Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10*, pages 103–110, New York, NY, USA, 2010. ACM.
- [36] E. Zhong, W. Fan, Y. Zhu, and Q. Yang. Modeling the dynamics of composite social networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13*, pages 937–945, New York, NY, USA, 2013. ACM.

Avaliação de características que influenciam nos votos de utilidade de opiniões sobre serviços em Português

Alternative Title: Assessement of features influencing the voting for opinions' helpfulness about services in Portuguese

Augusto C. S. Martins
Programa de Pós-Graduação em
Computação Aplicada
Universidade Tecnológica
Federal do Paraná - UTFPR
augustomartins@alunos.utfpr.edu.br

Cesar A. Tacla
Programa de Pós-Graduação em
Computação Aplicada
Universidade Tecnológica
Federal do Paraná - UTFPR
tacla@utfpr.edu.br

RESUMO

Este trabalho apresenta a aplicação de uma metodologia de avaliação de utilidade de opiniões com o objetivo de identificar quais características exercem maior influência na quantidade de votos de utilidade: básicas (ex. nota sobre o produto e/ou serviço, data da publicação), textuais (ex. tamanho das palavras, parágrafos) e semântica (ex. o significado das palavras do texto). A avaliação foi realizada em uma base de dados extraída do *site* TripAdvisor com opiniões sobre hotéis escritas em português. Resultados mostram que os usuários dão mais atenção a opiniões recentes com notas mais altas para valor e localização do hotel e com notas mais baixas para limpeza e qualidade dos quartos. Textos com valores pequenos para inteligibilidade (mais difíceis) recebem mais votos do que textos com valores grandes de inteligibilidade.

Palavras-Chave

Mineração de opiniões, Utilidade da opinião, Qualidade da opinião

ABSTRACT

This paper presents the application of a methodology for evaluation of usefulness of opinions with the aim of identifying which characteristics have more influence on the amount of votes: basic utility (e.g. ratings about the product and/or service, date of publication), textual (e.g. size of words, paragraphs) and semantics (e.g., the meaning of the words of the text). The evaluation was performed in a database extracted from TripAdvisor with opinions about hotels written in Portuguese. Results show that users give more attention to recent opinions with higher scores for value and location of the hotel and with lowest scores for cleanliness and quality of rooms. Texts with small values for intelligibility (more

difficult) receive more votes than texts with large values of intelligibility.

Categories and Subject Descriptors

H.3.1 [Information Systems]: Information Storage and Retrieval—*Content Analysis and Indexing*; I.1.2 [Computing methodologies]: Artificial intelligence—*Information extraction, Language resources*

General Terms

Experimentation

Keywords

Opinion mining, Opinion helpfulness, Opinion quality

1. INTRODUÇÃO

O grande número de opiniões geradas por usuários em *blogs*, redes sociais, *microblogs*, lojas virtuais, *sites* especializados em avaliações de produtos e serviços, fez o antigo "boca a boca" migrar para o mundo virtual, criando grandes comunidades eletrônicas. O conteúdo gerado por usuários não possui uma estrutura formal e tão pouco é simples de ser processado automaticamente. No entanto, como o número de opiniões cresce rapidamente, cada vez é maior o esforço do usuário para encontrar a informação desejada [22, 16]. Esse excesso de informação pode sobrecarregar os usuários. Além de numerosas, muitas das opiniões são fraudulentas, não-informativas, incompletas ou repetitivas. Além disso, opiniões úteis estão misturadas com um grande número de opiniões sem utilidade.

Uma alternativa utilizada por algumas empresas (i.e. Amazon.com, TripAdvisor, Booking.com) foi implementar um sistema de avaliação onde seus usuários informam se a opinião de outra pessoa foi útil ou não. Essa avaliação se dá através da pergunta "Esta opinião foi útil para você?", geralmente posicionada na sequência do conteúdo da opinião e somente com as opções "Sim" ou "Não" como resposta. Além disso, o total de votos positivos e o total geral de votos também podem ser apresentados, sendo estes utilizados no cálculo do índice de utilidade da opinião. Este índice é usualmente um dos critérios de ordenação das opiniões dos *sites* que utilizam esse sistema de votação [10].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

A tarefa mais importante quando se organiza um conjunto de opiniões é determinar quais serão úteis para os usuários. A maioria dos *sites* apresenta o conteúdo utilizando uma ordenação por data ou pontuação (i.e. nota ou números de estrelas atribuído pelo público) e também apresentam um resumo da avaliação de todos que votaram, como por exemplo "20 de 30 pessoas acham essa opinião útil" ou simplesmente apresentam o número de votos de utilidade que a opinião recebeu junto com o conteúdo do texto [11, 27]. No geral, as opiniões são criadas por pessoas que possuem necessidades, pontos de vista e tiveram experiências diferentes sobre um mesmo objeto e revelam quais atributos do produto ou serviço são os mais importantes. O sistema de votação de utilidade não resolve todos os problemas, pois a maior parte das opiniões não recebe voto e sem isso, o mesmo não cumpre o seu papel de indicar o conteúdo mais útil.

Portanto, a contribuição desta pesquisa é apresentar uma aplicação da metodologia descrita no trabalho de [5] em outra língua (português), com opiniões no domínio de serviços (hotelaria) e a utilização de métricas de inteligibilidade, a fim de identificar entre características básicas (ex. nota sobre o produto e/ou serviço, data da publicação), textuais (ex. tamanho das palavras, parágrafos), semântica (ex. o significado das palavras do texto) do conteúdo de opiniões de serviços de hotel, escritas em português, quais possuem maior influência nos usuários em sua avaliação de utilidade, a fim de definir um modelo capaz de selecionar um conjunto de opiniões que possuem maior probabilidade de receber votos.

O restante do artigo está organizado da seguinte maneira: a Seção 2 apresenta os trabalhos correlatos; a Seção 3 apresenta a metodologia, a seção 4, os experimentos realizados e os resultados obtidos; e as considerações finais na Seção 5.

2. TRABALHOS CORRELATOS

A revisão do estado da arte realizada a partir do estudo de [16], resultou na descoberta do trabalho de [5], através de buscas realizadas nas base de publicações científicas do *IEEE*, *ACM*, *Cite Seer X*, *Web Of Knowledge*, Periódicos Capes, com palavras-chave relativas a mineração de opiniões, análise da qualidade, utilidade da opinião e qualidade da opinião (e equivalentes em inglês¹). A partir dos resultados encontrados, foram selecionados trabalhos relacionados a qualidade da opinião no período de 2006 à 2014 ordenados por número de citações.

Na literatura, é possível encontrar trabalhos [11, 29, 9, 17, 18, 21, 28, 6, 19] que têm por objetivo estimar automaticamente a qualidade das opiniões através do número de votos de utilidade e outros atributos (ex. frequência de palavras, classes gramaticais) capturados por meio de mineração de dados e aprendizado de máquina, e baseadas no conteúdo das opiniões que receberam pelo menos um voto de utilidade.

Também há uma linha de pesquisa [26, 5, 10, 12, 15, 27] que tenta responder a pergunta: "quais os fatores que influenciam no número de votos que uma opinião recebe?". Esses trabalhos também utilizam técnicas de mineração de dados e aprendizado de máquina, porém consideram todas as opiniões e procuram analisar o quanto as propriedades básicas, textuais, erros de escrita, semântica, histórico de contribui-

ções de um mesmo autor e as opiniões fortemente polarizadas (i.e. muito positivas ou negativas, de acordo com as notas atribuídas pelos usuários) influenciam na leitura e avaliação das opiniões publicadas *online*.

Os artigos de [5] e [12] foram considerados pontos de partida para este trabalho, pois, o primeiro utiliza somente o conteúdo disponível de forma pública no *site* CNETD Download.com, sem a necessidade de informações restritas dos autores das opiniões, interação em redes sociais entre usuários e informações mais detalhadas das opiniões (i.e. análise de subjetividade, polaridade, etc.) e o segundo utiliza métricas de inteligibilidade da língua inglesa, a subjetividade das opiniões, os erros de escrita e o histórico de votos de utilidade que o autor da opinião recebeu, com o objetivo de determinar se essas características dos textos têm influência sobre o volume de vendas dos produtos analisados.

O trabalho de [12] afirma que testes de inteligibilidade são utilizados para quantificar diferentes tipos de texto dentro das áreas da ciência da informação. Em sua pesquisa, os autores aplicam quatro testes de inteligibilidade disponíveis para a língua inglesa: a) *Gunning's Fog* (FOG), b) *Flesh/Kincaid Reading* (FK), c) *Automated Readability-Index* (ARI), c) *Coleman-Liau Index* (CLI) e afirmam que as características textuais é que possuem maior influência na avaliação de utilidade das opiniões. Porém, em sua pesquisa [5], chega a conclusão que as características semânticas apresentam maior influência.

Além das conclusões diferentes, os autores utilizaram metodologias diferentes em dois pontos: (i) o primeiro não utiliza a Análise de Semântica Latente (*LSA*, do inglês *Latent Semantic Analysis* [13, 8]) e (ii) o segundo não utilizou métricas de inteligibilidade. Essa divergência cria uma lacuna para a investigação efetiva de quais características do texto são consideradas pelos usuários quando lêem e avaliam uma opinião como útil ou não. Contudo, as métricas de inteligibilidade da língua inglesa não se aplicam a língua portuguesa [4, 25]. Surge daí a pergunta: "quais os fatores que influenciam no número de votos que uma opinião escrita em português recebe?". Essa é a pergunta na qual se baseia este trabalho.

3. MÉTODO ADOTADO

Os experimentos foram realizados seguindo a sequência utilizada no trabalho de [5], porém adaptada para opiniões escritas em português e do domínio de serviços (hotéis). Os passos realizados foram: pré-processamento do texto para separar as características e realizar contagens textuais (ex. números de palavras, número de sentenças), a análise sintática, redução de termos através da *LSA* e a Análise de Fatores.

Dessa forma, este trabalho estende a pesquisa de [5] de três maneiras: (i) avaliar se é possível chegar a resultados aproximados, porém, utilizando opiniões no domínio de serviços de hotéis no lugar de *software* de computador (i.e. produtos, domínio utilizado em [5]); (ii) aplicar métricas de inteligibilidade utilizando uma implementação da ferramenta *Coh-Metrix-Port* de [25], para verificar se o nível de dificuldade de entendimento de textos influencia nos votos de utilidade; e (iii) utilizar a língua portuguesa² como idioma principal na avaliação das características textuais e semânticas das

¹*opinion mining, sentiment analysis, review quality, review helpfulness e opinion quality.*

²Esta pesquisa não distingue opiniões em Português do Brasil de outras variações da Língua Portuguesa.

opiniões.

3.1 Métricas de Inteligibilidade

Métricas de inteligibilidade, como as baseadas no número de sílabas por palavra e na média de palavras por sentenças em parágrafos, nem sempre traduzem a complexidade do texto. As técnicas de processamento de linguagem natural (PLN) possuem ferramentas capazes de verificar a inteligibilidade e automatizar a tarefa de classificar um texto de acordo com o nível de escolaridade ou audiência [24]. Existem somente duas ferramentas de avaliação da inteligibilidade para o Português, sendo estas, uma adaptação do índice *Flesh Reading Ease* (Martins et al. (1996) *apud* [25]) e uma nova abordagem de análise da inteligibilidade baseada na simplificação de textos para pessoas com baixo nível de escolaridade [25].

A ferramenta Coh-Metrix (Graesser et al. (2004), McNamara et al. (2002), Crossley et al. (2007) *apud* [25]), é utilizada para calcular os índices de coesão, coerência e a dificuldade de compreensão de um texto (em inglês) por meio de análise linguística de diferentes tipos: léxico, sintático, discursivo e conceitual. A coesão é definida por Graesser et al. (2003) *apud* [25] como “*características de um texto que, de alguma forma, ajudam o leitor a conectar mentalmente as idéias do texto*”. Já coerência é definida pelo mesmo autor como “*características do texto (ou seja, aspectos de coesão) que provavelmente contribuem para a coerência da representação mental*”.

Devido a incompatibilidade das métricas de inteligibilidade da língua inglesa com a língua portuguesa, foi necessário utilizar um conjunto de métricas adaptadas para o português [4], o Coh-Metrix-Port, desenvolvido por [25].

3.2 Coleta de dados

Os dados para esta pesquisa foram coletados de opiniões reais dos serviços de hotéis do *site* de viagens TripAdvisor, que indexa entre outras informações, as opiniões dos viajantes de diversas cidades em mais de 40 países em todo o mundo. As opiniões são exibidas em uma área pública do *site* que também possui informações do hotel (i.e. endereço, fotos, preços, avaliação geral), dispostas na forma de uma lista (ordenada da mais recente para a mais antiga) e limitada a 10 opiniões por página.

Aproveitando a escolha do Brasil para realização da Copa do Mundo de 2014, foi coletada e utilizada uma amostra aleatória com 10.920 opiniões de 2.457 hotéis das 12 cidades sede selecionadas para o evento do futebol: Rio de Janeiro (RJ), São Paulo (SP), Belo Horizonte (MG), Porto Alegre (RS), Brasília (DF), Cuiabá (MT), Curitiba (PR), Fortaleza (CE), Manaus (AM), Natal (RN), Recife (PE) e Salvador (BA). Apesar de domínios diferentes, em [5] foram utilizadas 3.460 opiniões e em [12], pouco mais de 36 mil opiniões. A quantidade de opiniões (absoluta e porcentual) que os hotéis de cada cidade da copa receberam é apresentada na tabela 1.

Conforme discutido no trabalho de [5], a maioria das opiniões não recebe voto de utilidade. O problema, que se repete em outros trabalhos citados anteriormente, também foi encontrado na distribuição de votos por opinião neste trabalho (Tabela 2). É possível notar que quando a quantidade de votos aumenta, o número de opiniões diminui expressivamente. A quantidade de opiniões que nunca recebeu voto ultrapassa 60% da base, enquanto a quantidade que recebeu

dez ou mais votos, não alcança 1% do total de opiniões.

3.2.1 Opinião dos usuários do TripAdvisor

A Figura 1 é um exemplo de uma opinião registrada no TripAdvisor. No canto superior esquerdo abaixo do título está a avaliação geral do hotel realizada pelo usuário, seguida pela data de publicação da opinião. Na sequência é apresentada o texto livre da opinião. A data da hospedagem no hotel e as avaliações das características do hotel, quando o usuário preenche esses dados, estão posicionados antes da pergunta sobre a utilidade da opinião (circulado). O TripAdvisor apresenta somente o número total de votos que a opinião recebeu, não havendo diferença entre votos positivos ou negativos.

3.2.2 Características básicas das opiniões

As características básicas podem ser extraídas de forma simples e sem a utilização de ferramentas de PLN (informações disponíveis na Figura 1). Foram utilizados: 1) o intervalo (em dias) entre a data da publicação da opinião e a data da viagem, 2) o extremismo da avaliação, valor estimado pela diferença entre a nota da opinião dada ao hotel e média geral de todos os usuários. Além desses itens, também foram utilizadas as notas em uma escala de 1 a 5 para as características: 3) custo-benefício, 4) localização, 5) qualidade do sono, 6) quartos, 7) limpeza e 8) atendimento.

Tabela 1: Quantidade de opiniões por cidade.

Cidade	Hotéis (Qtd.)	Opiniões (Qtd.)	%
Belo Horizonte	159	608	5.57
Brasília	87	673	6.16
Cuiabá	58	135	1.24
Curitiba	143	823	7.54
Fortaleza	219	903	8.27
Manaus	80	286	2.62
Natal	220	1038	9.51
Porto Alegre	111	536	4.91
Recife	98	552	5.05
Rio de Janeiro	529	2123	19.44
Salvador	246	831	7.61
São Paulo	507	2412	22.09
Total	2457	10920	100

3.2.3 Características textuais das opiniões

As características textuais representam o estilo de escrita do usuário e não podem ser extraídas sem um pré-processamento no conteúdo da opinião. Para este trabalho, foi implementada uma versão da ferramenta Coh-Metrix-Port, baseada na documentação disponível *online*³, somente com as contagens básicas (i.e. cálculo do inteligibilidade, número de sentenças, número de palavras, número de sílabas, média de palavras por sentenças, número de verbos, número de substantivos, número de adjetivos, número de advérbios e número de pronomes).

O índice *Flesh Reading Ease* avalia superficialmente a inteligibilidade de um texto. Esta métrica mede somente o número de palavras em sentenças e o número de letras ou sílabas por palavra sendo a única métrica que foi adaptada para o português [4]. A saída da fórmula *Flesh Reading*

³<http://www.nilc.icmc.usp.br:3000/index/info>

Ease é um número entre 0 e 100, e quanto mais alto o índice, mais fácil a leitura [4, 25]:

$$206.835 - (1.015 \times ASL) - (84.6 \times ASW) \quad (1)$$

onde, *ASL* é o tamanho médio de sentenças medido em número de palavras (o número de palavras dividido pelo número de sentenças) e *ASW* é número médio de sílabas por palavra (o número de sílabas dividido pelo número de palavras). Em português, a adaptação do *Flesch Reading Ease* resultou na fórmula:

$$248.835 - (1.015 \times ASL) - (84.6 \times ASW) \quad (2)$$

que corresponde à fórmula original salvo pela constante 248,835 que diferencia textos em inglês de textos em português⁴. De acordo com [25], os valores desse índice variam entre 100-75 (muito fácil), 75-50 (fácil), 50-25 (difícil) e 25-0 (muito difícil), que correspondem, respectivamente, às duas subdivisões da educação fundamental (1-4 e 5-8), ensino médio (9-11) e ensino superior. A partir dessa implementação simplificada do Coh-Metrix-Port, foi possível extrair as características textuais, tais como: 1) índice de inteligibilidade (em uma escala de 0 a 100) e as contagens de 2) sentenças, 3) palavras, 4) sílabas, 5) média de palavras por sentenças, 6) quantidade de verbos, 7) substantivos, 8) adjetivos, 9) advérbios e 10) pronomes.

Tabela 2: Quantidade de opiniões por votos de utilidade.

Votos de Utilidade	Opiniões (Qdt.)	%
0	6808	62.34
1	2479	22.7
2	882	8.08
3	340	3.11
4	154	1.41
5	73	0.67
6	35	0.32
7	42	0.38
8	15	0.14
9	17	0.16
≥ 10	75	0.69
Total	10920	100

3.2.4 Características semânticas das opiniões

As características semânticas representam a idéia principal da opinião e, segundo [5], causam grande influência no voto de utilidade da opinião. Para capturar a ideia principal das opiniões, foi utilizada a LSA, que é um método estatístico que explora a relação existente entre um conjunto de termos e os documentos nos quais estes aparecem, a fim de construir um padrão de relacionamento entre os termos e os documentos [7].

3.3 Pré-processamento das opiniões

Na etapa de pré-processamento, foram eliminadas todas as opiniões fora do intervalo de publicação definido entre 1º

⁴O número 42 é o valor de referência que, na média, diferencia textos em inglês e português citado por Martins et al. (1996) *apud* [25]

“Que lugar maravilhoso, super indico!”

●●●●● Avaliou em Abril 3, 2014

Que lugar maravilhoso gente!! Eu e meu marido fomos muito bem recebidos por todos que trabalham na pousada. Com isso o atendimento foi nota 10. A pousada está impecável, jardins bem cuidados, apartamentos limpos. A área da piscina é fantástica com seus ofurôs e banheiras com vista para as praias de Bombas e Bombinhas. A praia que fica bem perto da pousada é linda com água cristalina e areia branca. Nosso final de semana foi incrível. Indico à todos que se hospedem nesta linda pousada em Bombinhas - SC Abraços

se hospedou em Março 2014, viajou com a família

●●●●● Custo-benefício

●●●●● Localização

●●●●● Qualidade do sono

●●●●● Quartos

●●●●● Limpeza

●●●●● Atendimento

menos

Esta avaliação foi útil? Sim 11

Figura 1: Exemplo de opinião de uma pousada situada em Bombinhas/SC. Fonte: Captura de tela do site Tripadvisor.

de Janeiro e 1º de Agosto de 2014, com o intuito de abranger um período de tempo razoável antes e durante a Copa do Mundo de Futebol 2014 (realizada entre 12 de Junho e 13 de Julho/2014) e selecionar uma amostra aleatória com no máximo 10.920 opiniões⁵. Também foi calculado o intervalo (em dias) entre a data da publicação da opinião e a data da viagem e o nível de “intensidade”, estimado pela diferença entre a nota da opinião dada ao hotel e a nota média de todos os usuários. O processamento foi realizado utilizando a linguagem R [23] em um servidor Linux Debian 7, 64 bits com 120 GB de RAM, 2TB de disco e 24 núcleos (Intel Xeon E5-2420 1.90GHz).

3.4 Análise sintática e redução de termos

Na etapa de análise sintática, foram realizadas as contagens básicas da estrutura do texto (i.e. contagens de sentenças, palavras, sílabas, média de palavras por sentenças, quantidade de verbos, substantivos, adjetivos, advérbios e pronomes) e o cálculo do índice de inteligibilidade. Na sequência, foram feitas a remoção de *stopwords*⁶ e a redução das palavras ao seu radical com o objetivo de agrupar termos semelhantes, reduzindo a quantidade de palavras e a ambiguidade do texto para a próxima etapa. A ferramenta Apache OpenNLP [3] em conjunto com algoritmos próprios foi utilizada para realizar a classificação sintática e identificação de palavras e sentenças (i.e. processamento “*part-of-speech*”).

A filtragem de termos do texto resultou na construção da matriz termos-documentos. Essa matriz, foi construída utilizando a *TF-IDF* (do inglês, *Term Frequency-Inverse Document Frequency*). Esse processo consiste na separação do

⁵Essa quantidade de opiniões é o triplo do número de opiniões que [5] utiliza em seu trabalho.

⁶Palavras que podem ser consideradas irrelevantes. Exemplos: as, e, os, de, para, com, sem, foi, etc

conteúdo dos documentos (i.e. opiniões) em termos e a contagem de suas frequências. O resultado deste processamento gera uma matriz esparsa (conhecida com *Bag of Words* ou *BOW*) em que cada linha representa um documento e cada coluna um termo, isto é, seja d_i a linha e t_j a coluna da matriz, o elemento da matriz O_{ij} representa o número de vezes que o termo j aparece no documento i . Foram identificados 7.976 termos, resultando em uma matriz 10.920x7.976.

3.5 Análise semântica das opiniões

Na sequência da redução de termos e criação da matriz termos-documentos, [5] aplica a Decomposição de Valor Singular [14] (do inglês, *Singular Value Decomposition* ou SVD), sendo esta a etapa principal da LSA. Através desta técnica (SVD) são geradas três matrizes a partir da matriz termos-documentos original, e então uma nova matriz é obtida do produto destas três matrizes. A SVD é utilizada para localizar a informação semântica essencial em uma matriz de co-ocorrência das palavras. O objetivo é obter um espaço semântico condensado que representa a melhor relação entre os termos e documentos.

3.6 Análise Fatorial

A Análise Fatorial (do inglês, *Factor Analysis*) é um processo de mensuração que identifica variáveis similares em uma mesma estrutura. A denominação "Fatores SVD" é utilizada por [5] para definir as colunas da nova matriz de termos encontrada após à análise fatorial. Essa matriz resume ainda mais a matriz-semântica das opiniões e representa o conjunto de características semânticas que serão utilizadas nos modelos, isto é, um número reduzido de palavras que representam o conteúdo textual da opinião e que foram encontradas devido similaridades matemáticas dos termos, representando o nível mais simplificado da matriz termos-documento. A pesquisa de [5], utiliza conjuntos de 50, 100, 150 e 200 Fatores SVD, porém, o resultado apresentado para fins de demonstração utiliza somente 100 fatores, o mesmo número utilizado neste trabalho.

3.7 Definição dos modelos

A Regressão Logística Ordinal [20] [2], do inglês *Ordinal Logistic Regression* ou OLR, é utilizada por [5] para investigar a relação entre as características básicas, textuais e semânticas das opiniões e o número de votos de utilidade da opinião. A OLR é uma extensão do modelo de regressão logística binária, onde a variável dependente pode acomodar mais do que duas categorias de valores. A variável dependente deste estudo é um "ranking de utilidade" baseado no número de votos que uma opinião recebeu. O modelo OLR foi utilizado para determinar quais características implicam no aumento de votos de utilidade de uma opinião. As equações 3 e 4 definem a abordagem da OLR:

$$g(Pr(Y \leq i|x)) = \alpha_i + \beta'x, i = 1, 2, \dots, k \quad (3)$$

onde Y é a variável resposta (dependente) com k categorias codificadas em $1, 2, \dots, k$, $\alpha_1, \alpha_2, \dots, \alpha_k$, representa o parâmetro de localização para as k categorias da variável dependente, β é o vetor de coeficientes de regressão (declives), β' é o vetor β transposto e x é matriz das variáveis independentes. A função $g=g(\mu)$ é a função de ligação que permite (μ) assumir uma resposta linearmente relacionada as variáveis independentes. $Pr(Y \leq i|x)$ é a probabilidade de Y ser

Tabela 3: Descrição dos Modelos

Modelo	Descrição das variáveis	# variáveis
1	Básicas	8
2	Textuais	10
3	Semânticas (Fatores SVD)	100
4	Básicas + Textuais	18
5	Básicas + Textuais + Semânticas	118

menor ou igual a i , dado um valor de x . A escala de *odds* proporcionais foi usada como função de ligação, conforme definido na equação 4 [5]:

$$\begin{aligned} g_i(Pr(Y \leq i|x)) &= \ln \frac{P_r(Y \leq i|x)}{P_r(Y > i|x)} = \ln \frac{P_r \leq i|x}{1 - P_r(Y \leq i|x)} \\ &= \ln \frac{\odot_1(x) + \odot_2(x) + \dots + \odot_i(x)}{1 - (\odot_1(x) + \odot_2(x) + \dots + \odot_i(x))} \\ &= \alpha_i + \beta'x, \\ i &= 1, 2, \dots, k \end{aligned} \quad (4)$$

onde $\odot_i(x)$ é a probabilidade de se estar em uma classe i dado x . Nesta pesquisa, Y é o número de votos de utilidade das opiniões e estes foram agrupados de forma que os valores de "0" a "9" representam estes valores neste intervalo, enquanto "10" agrega "10 ou mais votos" em uma mesma categoria. A distribuição dos votos foi apresentada anteriormente na Tabela 2.

Para determinar como e quais características de uma opinião influenciam no número de votos de utilidade, foram criados cinco modelos OLR com várias combinações entre os três tipos de características básicas, textuais e semânticas seguindo a mesma combinação discutida no trabalho de [5]. A Tabela 3 apresenta a descrição dos modelos e a quantidade de variáveis independentes de cada um. Os três primeiros modelos (i.e. modelos 1, 2 e 3) serão utilizados para verificar a influência de cada característica separadamente. O modelo 4 congrega características básicas e textuais e o modelo 5 engloba os três tipos de características.

Alguns dos modelos possuem uma grande quantidade de variáveis, o que pode aumentar o ajuste do modelo e diminuir sua capacidade de previsão. Por isso é utilizado o método de seleção *stepwise*, que retira do modelo uma variável com maior *p-valor* e verifica-se se houve diminuição no ajuste do modelo. Caso não tenha ocorrido efeito no ajuste, a variável é retornada para o modelo e o mesmo processo é executado para a próxima variável. Esse processo é executado até que todas as variáveis tenham sido testadas ou que o modelo atinja seu ajuste mínimo. Para comparar os modelos, [5] utilizou três medidas de ajuste: taxa de classificação incorreta (do inglês, *misclassification rate*), o critério de informação de Akaike (i.e. *Akaike's information criterion*, ou AIC) e a razão de lift (i.e. *lift ratio*).

A taxa de classificação incorreta, é utilizada para calcular o quanto uma classificação pode ser considerada incorreta. É calculada através da proporção de classificações incorretas sobre o total de classificações do modelo e quanto maior a taxa de classificação incorreta do modelo, pior é o seu desempenho. Criado por Akaike [1], o critério de informação é utilizado para manter o compromisso entre a precisão do modelo e sua complexidade. Seu cálculo penaliza a inclu-

Tabela 4: Comparação dos Modelos

Modelo	Taxa de Classificação Incorreta	AIC	Lift ratio
1	0,62344	24651,57	0,98
2	0,37595	24602,58	1,18
3	0,37546	24712,84	0,88
4	0,37568	24588,56	1,16
5	0,37445	24530,49	1,27

são de muitas variáveis no modelo enquanto procura manter uma precisão razoável de maneira que, quanto menor o índice AIC, melhor o modelo. Esta métrica foi utilizada para comparar os cinco modelos criados para esta pesquisa. A razão de lift é utilizada para medir a performance de previsão de um modelo em um segmento aleatório da população. Como regra geral, quanto maior, melhor será o desempenho do modelo.

4. RESULTADOS OBTIDOS

Nesse estudo, chegou-se a uma conclusão semelhante a de [5] em que as características semânticas, ou seja, grupos de palavras encontradas no texto da opinião, possuem influências positivas e negativas no número de votos de utilidade que recebem. Complementar a essa conclusão, foi verificado que as pessoas tendem a dar mais atenção para opiniões mais recentes, que tenham notas mais altas em relação ao valor, localização do hotel e que apresentem notas mais baixas em relação à limpeza e qualidade dos quartos. Curiosamente, textos de opiniões com mais palavras, menos adjetivos e índices de inteligibilidade mais baixos (i.e. maior dificuldade) recebem mais votos do que opiniões sem estas características.

A Tabela 4 resume os resultados empíricos da comparação dos cinco modelos OLR. Assim como no trabalho de [5], o Modelo 5 (i.e. Básicas + Textuais + Fatores SVD), possui a menor taxa de classificação incorreta e os melhores índices AIC e *lift*, ou seja, obteve o melhor resultado entre os modelos. Esse resultado confirma que integrar as características semânticas junto com as características básicas e textuais, melhora o desempenho do modelo.

O Modelo 4 (i.e. Básicas + Textuais) possui o segundo melhor AIC, mas fica com a terceira posição em duas outras métricas (taxa de classificação incorreta e *lift*). O Modelo 2 (Textuais), com o terceiro melhor AIC, alterna entre a segunda e terceira posição para os índices *lift* e taxa de classificação incorreta. Os Modelos 1 (Básicas) e 3 (Fatores SVD), possuem os piores resultados em duas métricas (AIC e *lift*), o que pode indicar que características básicas e semânticas utilizadas isoladamente são os piores critérios para estimular o voto de utilidade.

A Tabela 5, mostra o resultado final do Modelo 5 com as variáveis selecionadas após a aplicação do método de seleção de variáveis da OLR, suas estimativas, desvio padrão, teste estatístico *Wald Chi-square* e o *p-valor*. Este modelo possui cinco variáveis básicas (intervalo, valor, quartos, limpeza, localização), cinco variáveis textuais (índice *Flesch Reading Ease*, número de palavras, número de substantivos, número de adjetivos e número de advérbios) e 47 variáveis semânticas (fatores SVD).

É importante notar que o *p-valor* de mais da metade das variáveis é menor que 0,05, o que indica que são estatisticamente significativas com 95% de nível de confiança. Porém, uma variável básica (número de palavras) e quatro variá-

veis semânticas⁷ tiveram o *p-valor* menor que 0,01, o que demonstra que são estatisticamente significativas com 99% de nível de confiança. A maioria dos *p-valores* do Modelo 5 está no intervalo de 0,05 e 0,01 de significância estatística com níveis de confiança entre 95% e 98% respectivamente.

Tabela 5: Estimativa dos parâmetros do Modelo OLR 5 (Básicas + Estilo + SVD)

Parâmetro	Coef.	Erro Padrão	Wald Chi-square	Pr > ChiSq
Intervalo	-0,0004	0,0002	6,72	0,0095
Valor	0,0512	0,0209	6,00	0,0143
Quartos	-0,0419	0,0223	3,53	0,0603
Limpeza	-0,0607	0,0236	6,59	0,0103
Localização	0,0305	0,0197	2,40	0,1210
Flesch R.E.	-0,0022	0,0010	5,33	0,0209
N. Palavras	0,0112	0,0024	22,78	< 0,0001
N. Substantivos	-0,0156	0,0072	4,72	0,0298
N. Adjetivos	-0,0164	0,0079	4,32	0,0376
N. Advérbios	-0,0213	0,0099	4,60	0,0320
SVD_4	0,1002	0,0405	6,12	0,0134
SVD_6	0,1356	0,0714	3,61	0,0575
SVD_9	0,4612	0,1864	6,12	0,0133
SVD_12	-0,3343	0,1294	6,68	0,0098
SVD_20	-0,2316	0,1553	2,23	0,1358
SVD_24	0,4922	0,3125	2,48	0,1152
SVD_21	0,7037	0,2110	11,12	0,0009
SVD_25	0,4151	0,2541	2,67	0,1024
SVD_33	0,0871	0,0592	2,16	0,1413
SVD_34	-0,4320	0,2860	2,28	0,1309
SVD_32	0,1636	0,0465	12,35	0,0004
SVD_35	-0,4477	0,2823	2,52	0,1127
SVD_36	0,6797	0,3603	3,56	0,0592
SVD_44	0,1556	0,0793	3,85	0,0498
SVD_43	0,6344	0,3186	3,96	0,0465
SVD_57	-0,0425	0,0238	3,19	0,0742
SVD_50	1,1196	0,5226	4,59	0,0322
SVD_77	-0,0878	0,0389	5,10	0,0240
SVD_58	0,2821	0,1207	5,46	0,0194
SVD_99	-0,0565	0,0306	3,40	0,0651
SVD_82	-0,3221	0,2178	2,19	0,1392
SVD_69	-0,0334	0,0188	3,15	0,0760
SVD_84	-0,0533	0,0254	4,41	0,0357
SVD_86	-0,3533	0,1852	3,64	0,0565
SVD_67	0,7409	0,2480	8,92	0,0028
SVD_83	-0,9946	0,5425	3,36	0,0668
SVD_100	1,0611	0,4500	5,56	0,0184

Ao contrário do trabalho de [5], que chega a conclusão de que palavras de até quatro letras em inglês, são mais simples de ler e podem significar mais votos para opinião, nesta pesquisa o resultado encontrado foi positivo para número de palavras e negativo para as outras variáveis textuais, ou seja, uma grande quantidade de palavras e um menor índice de inteligibilidade (i.e. mais difícil) com poucos substantivos, adjetivos e advérbios possui maior influência no número de opiniões. O intervalo de publicação da opinião e a data efetiva da utilização do serviço é negativo, indicando que quanto menor esse intervalo, maior a chance da opinião receber mais votos, isto é, opiniões incluídas pouco tempo após a hospedagem no hotel podem ser consideradas mais úteis devido ao fato da experiência com o serviço ser recente.

As opiniões "extremamente negativas" podem não exercer o mesmo efeito no domínio de serviços ou em outra cultura. Ao contrário do resultado encontrado em [5], que observou que as pessoas tendem a votar nas opiniões mais negativas, o Modelo 5 deste trabalho não incluiu a variável relacionada ao extremismo da opinião. A avaliação geral da nota pode ter dado lugar a avaliações das características específicas do hotel, conforme explicado a seguir. As variáveis contendo as notas sobre "valor", "quarto", "limpeza" e "localização" são

⁷SVD_12: "abertur"; SVD_21: "abram"; SVD_32: "abrind"; SVD_67: "aceit".

importantes para quantidade de votos de utilidade da opinião, ou seja, são as variáveis básicas que ajudam a diminuir o índice AIC sem prejudicar o desempenho do modelo. Por possuir uma estimativa positiva, os usuários tendem a votar em opiniões com notas mais altas para "valor" e "localização".

Apresentando estimativas negativas, valores mais baixos para "limpeza" e "quarto" atraem mais votos, o que indica tendência de procura por serviços com conforto e higiene. Estas quatro variáveis, podem indicar as características dos serviços de hotel que os usuários levam em consideração na escolha da empresa que irão contratar e estão diretamente ligadas ao domínio utilizado, isto é, certamente a utilização de opiniões sobre restaurantes terá outras variáveis básicas. Finalmente, os coeficientes de alguns fatores SVD (Seção 3.6) são positivos e outros negativos, o que indica que certas palavras possuem influência positiva, atraindo mais votos e outras palavras possuem o efeito contrário, desestimulando que o usuário avalie a opinião.

5. CONCLUSÃO

Neste trabalho foram realizados os experimentos descritos na pesquisa de [5] para encontrar quais características em um conjunto de opiniões pode influenciar a percepção de utilidade de uma opinião e, conseqüentemente, na quantidade de votos de utilidade que uma opinião pode receber. Assim como no trabalho de [5], o objetivo foi explorar se as características semânticas possuem influência no número de votos de utilidade. Foi possível observar nos resultados empíricos da pesquisa que essas as características semânticas efetivamente possuem impacto na avaliação de utilidade das opiniões também em língua portuguesa e no domínio de serviços da hotelaria. Curiosamente, a utilização do índice de inteligibilidade revelou que opiniões mais complexas e que possuem mais palavras são mais úteis do que opiniões com poucas palavras e mais fáceis de ler.

Estendendo o experimento de [5] e considerando algumas de suas limitações, foi possível adaptá-lo à mudança da Língua Inglesa para a Portuguesa e aplicá-lo em outro domínio (hotelaria). Assim como mencionado na conclusão de [5], as características semânticas são importantes para a percepção de utilidade das opiniões, mas quais são efetivamente as palavras que capturam a atenção dos leitores? Essa investigação pode levar a descoberta de padrões de textos que potencializam as chances de leitura e mais votos em uma determinada opinião.

As opiniões utilizadas são específicas para hotéis e possuem suas características próprias, como localização do hotel, detalhes referentes aos quartos, atendimento dos funcionários e outros, porém, a base de dados também possui informações sobre opiniões de Restaurantes e Pontos Turísticos das cidades escolhidas para a Copa do Mundo de 2014 e sendo possível comparar em um trabalho futuro se os resultados se mantêm, por exemplo, para opiniões sobre atrações turísticas. Além disso pode-se investigar a existência de um índice de inteligibilidade "recomendado" e que teria um efeito semelhante na quantidade de votos de utilidade. Outro trabalho futuro, seria retirar a limitação arbitrária de 100 Fatores SVD (utilizada no artigo de [5]) e investigar qual o conjunto mínimo de fatores que melhoram o desempenho do modelo penalizá-lo pela inclusão de um número excessivo de variáveis.

6. AGRADECIMENTOS

Os autores agradecem à Fundação Araucária, CAPES e CNPq pelo financiamento parcial deste trabalho.

7. REFERÊNCIAS

- [1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, Dec 1974.
- [2] J. A. Anderson. Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–30, 1984.
- [3] Apache. Opennlp, 2011.
- [4] E. M. F. Barboza and E. M. d. A. Nunes. A inteligibilidade dos websites governamentais brasileiros eo acesso para usuários com baixo nível de escolaridade. *Inclusão Social*, 2(2):19–33, 2007.
- [5] Q. Cao, W. Duan, and Q. Gan. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, 50(2):511–521, Jan. 2011.
- [6] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, page 141–150. ACM, 2009.
- [7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.
- [8] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.
- [9] A. Ghose and P. G. Ipeirotis. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *Proceedings of the ninth international conference on Electronic commerce*, page 303–310. ACM, 2007.
- [10] A. Ghose and P. G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512, Oct. 2011.
- [11] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, page 423–430. Association for Computational Linguistics, 2006.
- [12] N. Korfiatis, E. García-Bariocanal, and S. Sánchez-Alonso. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3):205–217, May 2012.
- [13] T. K. Landauer, P. W. Foltz, and D. Laham. *Introduction to Latent Semantic Analysis*, volume 25, pages 259–284. Discourse Processes, 1998.
- [14] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, 1998.

- [15] J. Lee. What makes people read an online review? the relative effects of posting time and helpfulness on review readership. *Cyberpsychology, Behavior, and Social Networking*, 16(7):529–535, July 2013.
- [16] B. Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.
- [17] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL*, page 334–342, 2007.
- [18] Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. pages 443–452. IEEE, Dec. 2008.
- [19] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, page 691–700. ACM, 2010.
- [20] P. McCullagh. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142, 1980.
- [21] M. P. O’Mahony and B. Smyth. Learning to recommend helpful hotel reviews. In *Proceedings of the third ACM conference on Recommender systems*, page 305–308. ACM, 2009.
- [22] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 115–124. Association for Computational Linguistics, 2005.
- [23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. ISBN 3-900051-07-0.
- [24] C. Scarton, C. Gasperin, and S. Aluisio. Revisiting the readability assessment of texts in portuguese. In *Advances in Artificial Intelligence-IBERAMIA 2010*, page 306–315. Springer, 2010.
- [25] C. E. Scarton and S. M. Aluísio. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61, 2010.
- [26] A. Talwar, R. Jurca, and B. Faltings. Understanding user behavior in online feedback reporting. In *Proceedings of the 8th ACM conference on Electronic commerce*, page 134–142. ACM, 2007.
- [27] J. Tang, H. Gao, X. Hu, and H. Liu. Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM conference on Recommender systems*, page 1–8. ACM, 2013.
- [28] O. Tsur and A. Rappoport. RevRank: a fully unsupervised algorithm for selecting the most helpful book reviews. In *ICWSM*, 2009.
- [29] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, page 51–57. ACM, 2006.

Avaliação de Técnicas de Mineração de Dados para Predição de Desligamentos em Sistemas Elétricos de Potência

Alternative Title: Data Mining Techniques Assessment for Shutdowns Prediction of Electric Power Systems

Anderson Trindade Maia Faculdade de Computação - ICEN - UFPA Rua Augusto Correa nº 01 – Guamá Belém – Pa - Brasil andersontrindade@ufpa.br	Jefferson Magalhães de Moraes Faculdade de Computação - ICEN - UFPA Rua Augusto Correa nº 01 – Guamá Belém – Pa – Brasil jmorais@ufpa.br	Yomara Pinheiro Pires Faculdade de Engenharia Elétrica - ITEC - UFPA Rua Augusto Correa nº 01 – Guamá Belém – Pa – Brasil yomara@ufpa.br
---	---	---

Aldebaro Barreto da Rocha Klautau Júnior Faculdade de Engenharia Elétrica - ITEC - UFPA Rua Augusto Correa nº 01 – Guamá Belém – Pa – Brasil aldebaro@ufpa.br	Daniel Martins Faculdade de Engenharia Elétrica - ITEC - UFPA Rua Augusto Correa nº 01 – Guamá Belém – Pa – Brasil odanielmartins@gmail.com
--	--

RESUMO

O volume de informações históricas relacionado ao comportamento dos sistemas elétricos de potência, fez aumentar significativamente a dimensão das bases de dados das empresas de energia, sem entretanto contribuir com a melhoria nos aspectos de operação, manutenção e qualidade do serviço, salvo para eventuais consultas do comportamento histórico de variáveis. Diversos trabalhos na literatura vêm sendo propostos para investigar diversos aspectos relacionados este problema. Alguns desses trabalhos abordam o problema da predição de ocorrência de desligamentos em sistemas de potência utilizando técnicas de mineração. No entanto, devido ao grande volume de dados, a busca por metodologias de preparação e seleção de dados torna-se importante. Neste trabalho, é apresentada uma proposta metodológica utilizando técnicas de mineração de dados aplicadas às séries temporais de medição fasorial sincronizada para auxílio da prevenção de desligamentos devido a desvios de tensão em sistemas elétricos de potência. São testadas algumas técnicas de mineração em dados obtidos em uma linha de transmissão de 230 kV que interliga as subestações de Tucuruí, Altamira, Rurópolis, constituindo o sistema elétrico Tramo Oeste do Pará.

Palavras-Chave: PMU's, Sistemas Elétricos de Potência, desligamentos e mineração de dados.

A permissão para fazer cópias digitais ou impressas de todo ou parte deste trabalho para uso pessoal ou de sala de aula é concedida sem custos desde que as cópias não sejam feitas ou distribuídas para o lucro ou vantagem e que as cópias possam suportar este aviso e a citação completa na primeira página. Para copiar o contrário ou republicar, ou ainda, postar em servidores bem como redistribuir as listas, será preciso autorização prévia específica e / ou uma taxa.

SBSI 2015, 26-29 maio de 2015, Goiânia, Goiás, Brasil.
Direitos autorais SBC 2015.

ABSTRACT

The volume of historical information related to the behavior of electric power systems, has significantly increased the size of the energy companies databases, but without contributing to the improvement in the aspects of operation, maintenance and quality of service, except for any queries historical behavior of variables. Several studies in the literature have been proposed to investigate various aspects this problem. Some of these works address the problem of disconnections of occurrence prediction in power systems using mining techniques. However, due to the large volume of data, the search for the preparation of methodologies and data selection becomes important. In this work, a methodology using data mining techniques applied the synchronized phasor measurement time series to aid the prevention of shutdowns due to voltage deviations in electrical power systems is presented. Are tested some mining techniques on data obtained in a 230 kV transmission line linking the substations of Tucuruí, Altamira, Rurópolis, constituting the electrical system Tramo western Pará.

Categories and Subject Descriptors

H.2.8 [Database Applications]
I.2.1 [Applications and Expert Systems]

General Terms

Algorithms

Keywords: PMUs, Electric Power Systems, shutdowns and data mining.

1. INTRODUÇÃO

O setor elétrico brasileiro ultimamente vem sofrendo um impacto mais acentuado para conseguir manter as necessidades de geração, transmissão e distribuição de energia elétrica (Figura 1) que possam atender de forma aceitável e efetiva o aumento da demanda de seus consumidores em energia elétrica.



Figura 1: Exemplo de Geração, Transmissão, Interligação e Distribuição em um Sistema Elétrico de Potência[1].

Como não bastasse ainda, as seguidas paralisações nas obras das hidrelétricas, necessárias ao atendimento da natural matriz energética do tipo hidráulica, vêm obrigando o governo a investir em energia térmica, de construção mais simples e rápida, porém, de agressão ambiental talvez superior à das hidráulicas devido à alta emissão de carbono na atmosfera aumentando ainda mais os problemas no setor elétrico.

O setor elétrico também vem apresentando um aumento da presença de pequenos produtores independentes, principalmente as Pequenas Centrais Hidrelétricas - PCHs e, portanto, sem controle centralizado, interligadas à rede elétrica de distribuição além do aumento na utilização de energia eólica, sabidamente instável e causadora de variações abruptas na geração vem trazendo à operação do Sistema Elétrico Interligado - SIN brasileiro um desafio sem precedentes.

Um outro fator que deve ser analisado dentro desse contexto, diz respeito à inevitável política de interligação dos diversos sistemas elétricos regionais em um único sistema elétrico de forma a criar uma colaboração das bacias hidrográficas. Este tipo de cenário vem apresentando uma maior propagação de distúrbios elétricos, antes localizados, agora atingindo diversas áreas ou regiões simultaneamente, aumentando muito o risco de grandes blecautes como os que se tem presenciado recentemente em algumas regiões do Brasil. Do ponto de vista de estabilidade de sistemas elétricos, as variações de tensão e frequência no SIN, vem se tornando cada vez mais frequentes, provocando a atuação de equipamentos de proteção e os consequentes desligamentos.

Diante deste cenário e procurando soluções para essas questões, as empresas do setor elétrico estão buscando utilizar tecnologias avançadas para aquisição e armazenamento de informações, as quais incluem os chamados distúrbios ou eventos de interesse[2]. Um típico exemplo dessa tecnologia são os equipamentos de oscilografia e, mais recentemente as PMUs¹ (Figura 2), que detectam e transmitem os valores de tensão, corrente e seus respectivos ângulos de fase. Essas informações são recebidas,

1 PMUs - Phasor Meter Units, são equipamentos de medição fasorial sincronizada que Esses equipamentos são capazes de amostrar grandezas elétricas sincronamente em subestações geograficamente afastadas, a taxas de amostragem de até 60 amostras por segundo [3].

organizadas temporalmente, armazenadas e disponibilizadas para uso pelo concentrador de medição fasorial PDC (Phasor Data Concentrator).

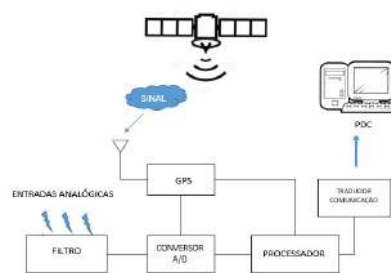


Figura 2 – PMU: Estrutura Básica

Os centros de controle de sistemas elétricos de potência (Figura 3), por sua vez, tratam e armazenam ordenadamente em banco de dados históricos, uma quantidade enorme de informação que, na maioria esmagadora das vezes, fica simplesmente armazenada e raramente observada no futuro. Entretanto, esta grande massa de informação, se sabe, guarda informações extremamente úteis e definitivamente necessárias para uma compreensão holística dos fenômenos. Se faz necessário, assim, minerar estes dados na busca por características, relações, associações, agrupamentos e etc., que tragam informações mais úteis sobre a operação do sistema elétrico de potência.



Figura 3 – Arquitetura Básica de um Sistema de Supervisão e Controle de um SEP.

Apesar da tecnologia avançada, minerar tais dados para inferir, por exemplo, relações de causa e efeito é ainda uma atividade incipiente [4]. Uma das razões é a falta de ações para relacionar os eventos com dados adicionais que poderiam ajudar a inferir a causa. Estabelecer esta relação é, hoje, uma das grandes preocupações nas indústrias de geração, transmissão e distribuição de energia elétrica.

Neste cenário cada vez mais competitivo, a tecnologia da informação, quando bem utilizada, torna-se um importante diferencial entre as empresas que primam por excelência na qualidade de serviço prestado. Neste contexto, surgem as aplicações e técnicas de mineração de dados cujo objetivo é descobrir padrões de comportamento e de novos conhecimentos sobre as informações armazenadas. Portanto, a gestão aleatória baseada principalmente em uma simples intuição vai dando lugar a inteligência do negócio.

Alguns trabalhos relacionados já abordaram metodologias para diagnóstico de desligamento em SEPs utilizando técnicas tais como interpolação em algoritmo simplista [5], Redes Neurais Artificiais (RNA) [6, 9], Lógica Fuzzy [7], Redes Bayesianas [8], *Random Forest*, *Support Vector Machine* (SVM) e *K-NearestNeighbor* (KNN) [1, 4].

Devido à dificuldade no tratamento dessas informações, ainda é difícil muitas das vezes adotar uma metodologia que seja capaz de mostrar com um tempo mínimo de antecedência a probabilidade de ocorrência de um evento de desligamento bem como também, verificar se este acontecimento está associado a fatores locais aparentes.

Em outra situação, por exemplo, pode não haver uma causa aparente, onde é comum apontar para uma falta transitória, como motivo de interrupção. Um exemplo típico é o desligamento por curto-circuito.

Após a ocorrência de um curto-circuito, se este não é permanente, o operador do sistema elétrico restabelece o sistema. Neste momento, o operador pode erroneamente interpretar esse evento (causa de interrupção) como uma ação de proteção. No entanto, na realidade, o curto-circuito e a atuação de proteção podem ser apenas consequências de algum evento anterior que ocorreu na rede, tais como condutores esgarçados, isoladores danificados ou poluídos, e muitos outros fatores que podem interromper o fornecimento de energia. A utilização correta da metodologia a ser adotada para preparação dos dados a serem minerados, é tão importante quanto os resultados da mineração no conhecimento do fenômeno, pois muitas vezes permite identificar a recorrência de um conjunto de desligamentos que podem estar originando a degradação da rede.

Com base nesse contexto, o objetivo deste artigo sinaliza para a necessidade de se investigar alternativas que permitam criar uma metodologia cujo objetivo seja a preparação dos dados das informações históricas de sistemas elétricos utilizando um processo de identificação, transformação, seleção e validação de informações para posterior mineração de dados em base de dados de sequência de eventos (SOE – *SequentialOfEvents*) e Medição Fasorial Sincronizada de forma a auxiliar os especialistas na obtenção de diagnóstico e especificamente na prevenção de desligamentos em sistemas elétricos de Potência.

2. METODOLOGIA

2.1 Área de Estudo

A área de estudo e investigação da pesquisa compreende os dados obtidos da medição fasorial sincronizada correspondente as subestações de Tucuruí, Altamira e Rurópolis que estão interligadas por uma linha de transmissão de 230 kV. Este trecho do Sistema Interligado Nacional – SIN, é denominado Tramoeste do Pará conforme ilustra a Figura 4.



Figura 4 – Região do Tramoeste.

Considerando que o fenômeno em observação é do tipo elétrico em alta tensão (230kV), alguns procedimentos devem ser adotados para extrair as informações do sistema elétrico.

2.2 Metodologia Proposta

A metodologia adotada trata inicialmente das informações provenientes das PMUs que não foram projetadas para um fim específico, mas que apresentam em seus dados eventos de desligamento. Assim, durante a fase de pré-processamento, apesar do PDC oferecer formas de onda de tensão e corrente (fases A, B e C), em sistemas de elevada tensão, como é o caso da aplicação desta pesquisa, as três fases são equilibradas, isto é, seus valores RMS (*Root Mean Squared*) são praticamente equivalentes, e as defasagens entre os fasores se mantêm razoavelmente próxima a 120 graus. Desta forma, foi utilizada apenas uma das fases, como por exemplo a fase B.

No caso do problema da predição do evento de desligamento, as instâncias (exemplos) são compostas por um conjunto de informações reais de maior relevância presentes na base para verificação da ocorrência ou não do evento, cujo resultado corresponde à classe dessas instâncias. Os campos selecionados para aplicação da metodologia são apresentados de forma simplificada na Tabela 1:

Tabela 1 – Campos da base de dados utilizados para mineração.

ID_Variável	Data_hora	Valor Medido
-------------	-----------	--------------

Onde cada campo é descrito como segue:

ID_Variável: Número de identificação da variável.

Data_hora: Valores de tensão da fase B medidos pelos PMUs de Tucuruí na barra 2, de Rurópolis na barra 1 e pelo de Altamira na barra 1.

DIF_FASE: Valores da diferença de fase entre a fase B medida pelo PMUs de Tucuruí na barra 2 e a fase B medida pelo PMU de Rurópolis na barra 1. De forma análoga, obteve-se também os Valores da diferença de fase entre a fase B medida pelo PMUs de Tucuruí na barra 2 e a fase B medida pelo PMU de Altamira na barra 1.

Vale ressaltar, que os campos TENSÃO e DIF_FASE são um conjunto de valores reais fornecidos pela Eletrobras-Eletronorte. Esses valores são medições realizadas continuamente no decorrer do mês de agosto de 2013. Este período foi escolhido, pois apresentou uma incidência maior de desligamentos quando comparados a outros meses do mesmo ano. Diante desse cenário e com objetivo de gerar um arquivo histórico passível de ser analisado por um especialista, os campos TENSÃO e DIF_FASE, foram divididos em vários subcampos, sendo um subcampo para cada medição pertencente a um determinado intervalo de tempo.

Portanto, para aplicar as técnicas de mineração de dados na tarefa de predição de ocorrências de desligamento em SEPs, foi composta uma base a partir dos valores RMS da tensão da fase extraídas de uma PMU, considerando uma janela de tempo de 60 segundos antes da ocorrência do evento. Assim, os dados coletados resultaram em uma base com 392 instâncias (exemplos), sendo 77% registros relacionados a eventos de desligamento e 22% de registros de não desligamento. A Figura 5 apresenta as etapas desenvolvidas desde o processo de transformação até a

aplicação das técnicas de mineração de dados nas medições fasoriais coletadas.

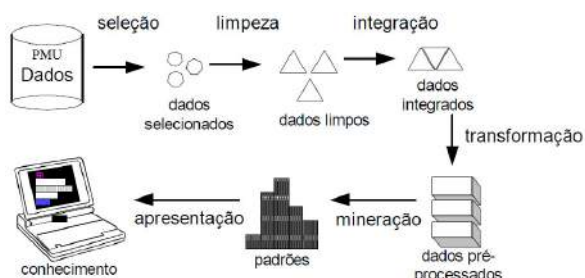


Figura 5 – Etapas do processo de Mineração (Ribeiro, 2004).

Três scripts (BinaryReader.java, CsvFormat.java, e AlarmFormat.java) foram desenvolvidos para etapa de seleção e pré-processamento dos dados. Já as etapas de transformação, os dados extraídos dos equipamentos são em formato binário. E na etapa de mineração dos dados foi utilizado o software WEKA(WaikatoEnvironmentKnowledgeAnalysis) que será mais detalhado nas próximas seções.

- BinaryReader.java: foi o primeiro script implementado, este arquivo realiza a leitura de arquivos binários provenientes da PMU. Estes arquivos contêm as medições fasoriais das estações de Rurópolis, Altamira e Tucuruí.
- CsvFormat.java: Este script realiza a etapa de pré-processamento dos dados selecionados. O arquivo de entrada é um arquivo no formato .csv, exportado pelo script BinaryReader.java, e o arquivo de saída é um arquivo no formato .arff. O arquivo de saída do CsvFormat já está preparado para posterior transformação e mineração por meio da ferramenta WEKA.

O algoritmo possui um atributo atualizável via código que determina o limite T de tempo de até T segundos antes do desligamento que será possível de se coletar os dados. Outro atributo importante do algoritmo é o windowSize, que define a janela de tempo que irá ser utilizada pelo software de mineração de dados para tentar classificar determinado evento. Um exemplo segue na Figura 6.



Figura 6 – Gráfico da tensão de uma estação durante o momento de um desligamento.

O arquivo de saída .arff (Attribute-Relation File Format) do CsvFormat.java está organizado da seguinte forma:

- Um cabeçalho – que contém informações sobre quais são os atributos deste conjunto de dados.
- Instâncias – que possui várias linhas, cada linha representa uma instância, com seus valores para cada atributo.

Esta é a organização aceita pelo programa de mineração WEKA. De cada janela de análise é sintetizada uma instância e cada medição fasorial de uma determinada janela compõe um campo de atributo de uma instância. A quantidade de atributos das instâncias depende do tamanho da janela de análise windowSize. Se uma janela de 60 segundos há 500 amostras, cada instância irá possuir 500 atributos.

No caso da mineração efetuada nesta pesquisa, cada instância possui 100 atributos. Estes atributos são nomeados r0,r1,r2,...,r99. Os primeiros 50 atributos (r0 à r49) rereferenciam 50 amostras de tensão igualmente espaçadas no tempo, enquanto que as últimas 50 amostras (r50 à r99) referenciam as amostras de fase medidas ao mesmo tempo que as amostras de tensão. Essas amostras são coletadas por uma PMU de uma determinada estação. A taxa de amostragem de uma PMU, na época da medição, é de 10 Hz, ou seja, 10 amostras por Segundo.

Como um conjunto muito grande de atributos pode prejudicar a mineração de dados, foi realizada uma subamostragem das amostras coletadas pela PMU, uma variável chamada samplePerInstance foi criada para armazenar o valor da subamostragem, seu valor foi configurado para 50 (50 amostras de tensão e 50 amostras de fase).

A partir das variáveis samplesPerInstance, samplingFrequency(10Hz) e windowSize (tamanho da janela em segundos) pode-se criar um fator de subamostragem definido pela formula:

$$X = \frac{WS * SF}{SP}$$

Onde,

WS = Tamanho da Janela

SF = Frequência de amostragem do PMU

SP = Quantidade de atributos de cada Instância.

A cada x amostras, uma amostra é copiada para o arquivo no formato aceito pelo weka e o restante é descartado.

As etapas (Figura 7) seguintes tratam efetivamente da avaliação sistemática dos classificadores adotados para prevenção de desligamentos.

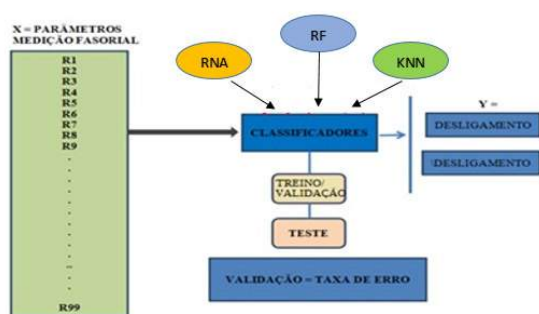


Figura 7 – Modelo de classificação adotado.

3. CLASSIFICADORES

Um *classificador* F é um mapeamento $F : R^k \rightarrow (1, \dots, Y)$, onde K é a dimensão do vetor de entrada $z \in R^k$ e o rótulo $y \in (1, \dots, Y)$ é a classe. Quando se treina um classificador usando aprendizado supervisionado, é dado um conjunto de treinamento $T = \{(z_1, y_1), \dots, (z_v, y_v)\}$ contendo V exemplos de (z, y) [10].

Um conjunto de teste $\{(z_1, y_1), \dots, (z_r, y_r)\}$ contendo R exemplos e disjuncto do conjunto de treino pode ser usado para calcular a taxa de erro de classificação

$$E_f = \frac{1}{R} \sum_{i=1}^R \tau(F(z_i) \neq y_i)$$

onde T é a função indicador, que é um (1) caso o argumento seja verdadeiro e zero (0) caso contrário. O erro E_f é uma estimativa da capacidade de generalização do classificador [11].

Para classificação este trabalho utilizou os algoritmos do software WEKA (*Waikato Environment for Knowledge Analysis*), software livre de domínio público assistido pela licença GPL (*General Public License*), e desenvolvido na Universidade de Waikato na Nova Zelândia, sendo disponibilizado em <http://www.cs.waikato.ac.nz/ml/weka/>. Assim, as próximas seções discutirão brevemente este software e as seções seguintes listarão os principais classificadores utilizados.

3.1 Weka

O WEKA é um software de referência reconhecido pela comunidade científica em aprendizado de máquina e mineração de dados [11-13]. O sistema é formado por um conjunto de algoritmos de aprendizado de máquina, implementado em linguagem Java, que o torna acessível nos principais sistemas operacionais.

O WEKA inclui algoritmos de regressão, classificação, agrupamento, regras de associação e seleção de parâmetros (atributos). Atualmente está na versão estável 3.6 que usa o formato de arquivo com extensão *arff* [10]. O mesmo possui diversos algoritmos de Inteligência Computacional de classificação e este trabalho usa apenas alguns dos principais, tais como Rede Neural Artificial (RNA) multicamadas treinadas com algoritmo *backpropagation*, Random Forest (RF) e K-nearestneighbor (KNN).

3.2 Redes Neurais Artificiais

As RNA são sistemas paralelamente divididos em camadas interligadas por pesos sinápticos e compostas por neurônios artificiais que calculam uma certa função matemática (usualmente não-linear). O algoritmo utilizado neste trabalho para treinamento da RNA foi o *backpropagation* também chamado de regra delta generalizada. Nesta pesquisa foi utilizada a classe *MultilayerPerceptron* que é o padrão do WEKA e os principais parâmetros dessa implementação são: (-L) - corresponde à taxa de aprendizagem utilizada pelo algoritmo *backpropagation* e este valor deve ser entre 0 e 1 (Padrão é 0,3); (-M) - taxa de momento para o algoritmo *backpropagation* e este valor deve ser entre 0 e 1 (Padrão é 0,2); (-N) - este parâmetro corresponde ao número de épocas para treinamento da rede e o padrão é 500; (-H) - corresponde à quantidade de camadas ocultas que podem ser criadas na rede [10,14].

3.3 Random Forest

O RF é a referência algorítmica em árvores de decisão. A ideia desse algoritmo é dividir o espaço de instâncias em subespaços, no qual os mesmos são ajustados em diferentes modelos, sendo formalmente estruturado em um grafo acíclico em que cada nó, ou é um nó de divisão com dois ou mais sucessores dotado de um teste condicional, ou um nó folha rotulado com uma nova classe. No WEKA o RF tem inúmeros parâmetros, sendo os principais são: (-depth) - responsável pela profundidade da árvore; (K) - número de recursos em cada iteração; (-I) - número de árvores do modelo [10,15].

3.4 K-nearest neighbor

O classificador KNN utiliza os próprios dados de treinamento como modelo de classificação, isto é, para cada novo padrão que se quer classificar, usa os dados do treinamento para verificar quais são os exemplos nessa base de dados que são mais próximos do padrão em análise, ou seja, baseado nas distâncias. A cada novo padrão a ser classificado faz-se uma varredura nos dados de treinamento, o que provoca um grande esforço computacional. O KNN no WEKA é implementado na classe *IBK* e seus principais parâmetros são: (-N) que é igual ao número de centros (ou K) e (-S) que gera aleatoriamente os centros [10,14].

4. RESULTADOS

As métricas utilizadas para avaliar o desempenho dos classificadores na prevenção de desligamentos em sistemas elétricos de potência, levou em consideração a variação dos valores nos parâmetros dos classificadores, assim como a acurácia, taxa de erro, taxa de especificidade, taxa de sensibilidade e medida F . Sendo que nas seções posteriores é apresentado os resultados do estudo com mais detalhes.

4.1 Resultado da Seleção do Modelo

Os valores que definem os modelos dos classificadores, tais como o número de neurônios na camada escondida de uma rede neural, por exemplo, foram escolhidos por meio de um procedimento de seleção automática de modelo. A Tabela 2 apresenta o conjunto de valores testados (*grid*) adotado nesse procedimento. É importante destacar que os valores escolhidos para compor o *grid* foram baseados em experimentos anteriores. Para validação dos

resultados foi adotado validação-cruzada (cross-validation) considerando 10-folds[11].

Tabela 2: Grid de seleção do modelo.

Classificador	Parâmetros	ValoresGrid	Melhor valor
RNA	H	3, 5, 7, 9 e 11	5
RF	I	5, 10, 15, 20 e 25	10
KNN	K	1, 3, 5, 7 e 9	7

Como pode-se observar na Tabela 2 em Valores Grid, foram feitos cinco experimentos para cada classificador com valores diferentes do parâmetro apresentado. A Rede Neural obteve uma melhor taxa de classificação com cinco neurônios em uma camada escondida (H), no qual foram feitos experimentos com 3, 5, 7, 9 e 11 neurônios respectivamente. Random Forest utilizou uma faixa de 5 a 25 árvores aleatórias no parâmetro (I), sendo que o valor ótimo foi com 10 árvores. E por fim o KNN no parâmetro K obteve uma melhor performance de classificação com valor de sete vizinhos mais próximos, sofrendo uma variação entre 1 e 9 dos cinco experimentos.

4.2 Resultado da Acurácia, Taxa de Erro, Taxa de Especificidade, Taxa de Sensibilidade e Medida F

Outros elementos relevante para mensurar o desempenho do classificador é a acurácia (taxa de acerto) e taxa de erro. Alguns classificadores são capazes de prover scores de confiança $f_i(x)$ para cada classe $i = 1, \dots, Y$, tais como a probabilidade de distribuição sobre y . Por conveniência, pode-se assumir que todo classificador retorna um vetor $(f_1(x), \dots, f_Y(x))$ com Y scores. Se o classificador naturalmente não retornar scores de confiança, o vetor de scores é criado com um score unitário $f_j(x)=1$ para a classe j sugerida pelo classificador, enquanto que os scores das outras classes são zero: $f_i(x) = 0, i \neq j$. Assim, a decisão final é sempre baseada no valor máximo dos scores (chamada regra *max-wins*).

Um conjunto de teste $\{(x_1, y_1, \dots, x_R, y_R)\}$ contendo R exemplos e disjunto do conjunto de treino pode ser usado para calcular a taxa de erro de classificação (Equação 4).

$$E_f = \frac{1}{R} \sum_{i=1}^R \tau(F(x_i) \neq y_i)$$

onde T é a função indicador, que é um (1) caso o argumento seja verdadeiro e zero (0) caso contrário. O erro E_f é uma estimativa da capacidade de generalização do classificador [11,12].

A taxa de erro varia entre 0 e 100, e valores próximos a 0 são considerados os melhores índices nas previsões dos

classificadores. Em [16] afirma-se que a acurácia (taxa de acerto), conforme a Equação 5, serve como complemento da taxa de erro.

$$ac(f) = 1 - E_f$$

[17,18] colocam que a taxa de acerto também varia entre 0 a 100, entretanto ao contrário da taxa de erro, valores próximos a 100 apresentam melhores previsões nas classificações. Foram adotadas também outras medidas de desempenho na previsão dos eventos de desligamentos, como por exemplo: a taxa de precisão (P), a taxa de sensibilidade (T_s) e a taxa de especificidade (T_e) e a medida F . O autor também chama atenção que tanto a acurácia como a taxa de acerto, assim como a P , T_s , e T_e e a medida F podem ser extraídas da matriz de confusão, pois as linhas da mesma contem valores² que representam as classes verdadeiras, e as colunas, as classes preditas pelo classificador (Equações 6, 7, 8 e 9).

$$P(f) = \frac{VP}{VP + FP}$$

$$T_s = \frac{VP}{VP + FN}$$

$$T_e = \frac{VN}{VN + FP}$$

$$medida - F(f) = 2 \left[\frac{precisao(f) * sens(f)}{precisao(f) + sens(f)} \right]$$

A Tabela 2 apresenta os resultados dos classificadores utilizados no estudo com as métricas adotadas na previsão dos eventos de desligamentos no sistema TRAMOESTE.

Tabela 2: Resultado da Taxa de acerto (acurácia), Taxa de erro, Taxa de sensibilidade e Taxa de especificidade, Medida F dos classificadores utilizados no estudo

Classificador	Acurácia(%)	Taxa de Erro(%)	Ts(%)	Te(%)	MedidaF
RNA	87,50	12,50	88,52	91,12	90,35
RF	81,25	18,50	84,30	86,75	85,82

(4)

2 FN: Falsos negativos: valor de exemplos classificados incorretamente como negativos; FP: Falsos positivos: exemplos classificados corretamente como negativos; VN: Verdadeiros negativos: exemplos classificados como verdadeiros incorretamente e VP: Verdadeiros Positivos: correspondem a todos a todos os exemplos classificados corretamente [17].

KNN	84,37	15,63	86,34	88,70	87,11
-----	-------	-------	-------	-------	-------

De acordo com a Tabela 2 observa-se que a RNA obteve o melhor desempenho em relação aos demais classificadores com acurácia e taxa de erro com valores de 87,2% e 12,2% respectivamente. Enquanto que o KNN foi o classificador com menor desempenho com 84,37% e 81,25% respectivamente na mesma métrica.

5. CONSIDERAÇÕES FINAIS

O estudo utilizou uma metodologia baseada em algoritmos de Inteligência Computacional, como Redes Neurais Artificiais, KNN e Random Forest, auxiliado por mecanismos de seleção automática de modelo na predição de eventos de distúrbios em linhas de transmissão de energia elétrica na rede Tramoeste como prevenção de desligamentos em sistemas elétricos de potência.

As métricas adotadas de acurácia (taxas de acerto), taxa de erro, taxa de especificidade, taxa de sensibilidade e medida F foram utilizadas para validar os resultados obtidos. Entre as técnicas computacionais utilizadas, a Rede Neural Artificial foi a que teve um desempenho mais satisfatório na predição de eventos de distúrbios em linhas de transmissão de energia elétrica na rede Tramoeste.

Sendo que a grande variação sazonal climática na área de estudo podem provocar distúrbios elétricos, assim como outras causas dessas alterações no sistema elétrico interfere em resultados mais precisos. Essa precisão também é afetada com a limitação de recursos estruturais e humanos.

Portanto, a aplicação de técnicas de inteligência computacional nas predições de eventos de distúrbios em linhas de transmissão de energia elétrica na rede Tramoeste como prevenção de desligamentos em Sistemas Elétricos de Potência pode se transformar em uma contribuição objetiva na administração dos recursos elétricos. Possibilitando pesquisas futuras que possam utilizar outras técnicas, além de Redes Neurais Artificiais, KNN e Random Forest para auxílio de melhor gerenciamento e prevenção em sinistros do sistema elétrico.

AGRADECIMENTOS

Os autores agradecem a Eletrobras-Eletronorte por ter fornecido as bases de dados e pelo suporte financeiro.

REFERÊNCIAS

- [1] Morais, J. M. (2011). Avaliação de Desempenho de Classificadores de Faltas em Sistemas Elétricos de Potência. Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica. Universidade Federal do Pará.
- [2] Bollen, M., (2009). Bridging the gap between signal and power. IEEE Signal Processing Magazine, 26:12, July.
- [3] Ree, J. D. L., Centeno, V., Thorp, J. S., and Phadke, A. G. (2010). Synchronized phasor measurement applications in power systems. IEEE Trans. Smart Grid, 1(1):20–27.

- [4] Morais, J. M., Pires, Y.P., Cardoso, C., Klautau, A. B. R. (2010). A Framework for Evaluating Automatic Classification of Underlying Causes of Disturbances and Its Application to Short-Circuit Faults. IEEE Transactions on Power Delivery, v. 25, p. 2083-2094.
- [5] Krishnanand, K. R., Dash, P. K., Naeem, M. H. (2015). Detection, classification, and location of faults in power transmission lines. Journal of Eletrica Power & Energy System, v. 67, p. 76-86.
- [6] He, Z., Lin, S., Deng, Y., Li, X., Qian, Q. (2014). A rough membership neural network approach for fault classification in transmission lines. Journal of Eletrica Power & Energy System, v. 61, p. 429-439.
- [7] Liu, Y., Schulz, N.N. (2002). Intelligent system applications in distribution outage management. Power Engineering Society Winter Meeting. IEEE, Vol. 2, pp. 833 – 837.
- [8] Tronchoni, A. B. (2008). Identificação de Causas de Desligamentos Não Programados em Redes de Distribuição. Dissertação de Mestrado, PPGE, PUCRS.
- [9] Pretto, C. O., Rancich, G.V., Lemos, F.A.B., Rosa, M.A. (2003). Forced Outages Information Treatment System and Cause Identification Based on Mobile Computing and Neural Networks. IEEE Bolonha Power Tech, Bolonha, pp. 1-6.
- [10] Arouche Freire, J. C., Santos, Y. S. R. N., Morais, J. M., Oliveira, T. F., Hauser-Davis, R. A., Saraiva, A. C. F. (2014b). Avaliação de Desempenho de Classificadores de Ciclos Hidrológicos em Reservatórios de Água na Região Amazônica. In: XXXIV Congresso da Sociedade Brasileira de Computação, WCAMA, Brasília-DF.
- [11] Witten, I. H., Frank, E. (2005). Data mining: practical machine learning tools and techniques with java implementations. Edition Morgan Kaufmann.
- [12] Hall, M.; Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The weka data mining software: an update. SIGKDD Explor. Newsl. 11, 10-18.
- [13] Guandaline, V. H., Merschmann, L. H. de C. 1. (2013). Avaliação de Técnicas de mineração de dados para predição de ocorrência do mexilhão. TrilhasTécnicas - SBSI, 2014.
- [14] Haykin, S. (1998). Neural Networks: A comprehensive Foundation. Prentice Hall, 2 edition, July 1998. p. 900.
- [15] Zhou, Q., Zhou, H., Zhou, Q., Yang, F., and Luo, L. (2014). Structure damage detection based on random forest recursive feature elimination. Mechanical Systems and Signal Processing.

- [16] Faceli, k., Lorena, A. C., Gama, J., Carvalho, A. C. P. L. F. *Inteligência artificial: uma abordagem de aprendizado de máquina*. Editora LTC, 2011.
- [17] Arouche Freire, J. C., Lobato, T. C., Moraes, J. M., Oliveira, T. F., Hauser-Davis, R. A., Saraiva, A. C. F. (2014a). Support Vector Machine and Artificial Neural Networks for Hydrological Cycles Classifications of a Water Reservoir in the Amazon. *Journal of Communication and Computer*, v. 11, p. 111-117.
- [18] Riflin, R., Klautau, A. (2004). In defense of one-vs-all classification. *Machine Learning Research* 5, 101-141.

Um Método Não Supervisionado para o Povoamento de Ontologias a partir de Fontes Textuais na Web

Alternative Title: An Unsupervised Method for Ontology Population from Textual Sources on the Web

Fábio Lima
Universidade Federal da
Bahia - UFBA
Rua Ademar de Barros -
Ondina
Salvador, Bahia
fabio.lima@dcc.ufba.br

Hilário Oliveira
Universidade Federal de
Pernambuco - UFPE
Av. Professor Moraes Rego,
1235 - Cidade Universitária
Recife, Pernambuco
htao@cin.ufpe.br

Laís Salvador
Universidade Federal da
Bahia - UFBA
Fraunhofer Project Center for
Software and Systems
Engineering
Salvador, Bahia
laisns@ufba.br

RESUMO

O crescimento na produção e disponibilização de informações não estruturadas na Web aumenta diariamente. Essa abundância de informações desestruturadas representa um grande desafio para a aquisição de conhecimento que seja processado por seres humanos e também por máquinas. Nesse sentido, ao longo dos anos diversas abordagens têm sido propostas para a extração automática de informações a partir de textos escritos em linguagem natural. Contudo, ainda existem poucos estudos que investigam a extração de informações a partir de textos escritos em português. Diante disso, o objetivo deste trabalho é propor e avaliar um método não supervisionado para o povoamento de ontologias utilizando a Web como grande fonte de informações, no contexto da língua Portuguesa. Os resultados obtidos com os experimentos realizados foram encorajadores e demonstraram que a abordagem proposta obteve uma taxa de precisão média de 67% na extração de instâncias de classes ontológicas.

Palavras-Chave

Ontologias, Povoamento de Ontologias, Extração de Informações

ABSTRACT

The increasing in the production and availability of unstructured information on the Web grows daily. This abundance of unstructured information is a great challenge for acquisition of structured knowledge. Many approaches have been proposed for extracting information from texts written in natural language. However, only a few studies have investigated the extraction of information from texts written in

Portuguese. Thus, this work aims to propose and evaluate an unsupervised method for ontology population using the Web as a big source of information in the context of the Portuguese language. The results of the experiments are encouraging and demonstrated that the proposed approach reached a precision rate of 67% in the instances of ontological classes extraction.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Information filtering; I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods—Representation languages

General Terms

Measures, Method

Keywords

Ontologies, Ontology Population, Information Extraction

1. INTRODUÇÃO

Com o crescimento na produção de informações em formato digital disponibilizadas na Web é constante o interesse no desenvolvimento de técnicas automáticas capazes de recuperar, analisar e sumarizar esse grande volume de informações. Atualmente, a Web pode ser considerada como o maior repositório de informações do mundo, contemplando os mais variados domínios do conhecimento. Por exemplo, a biomedicina possui uma vasta literatura contendo informações sobre novas doenças e seus tratamentos, sintomas, microrganismos causadores de enfermidades, dentre outras informações importantes. De maneira similar, outros domínios, como o de notícias, possuem informações sobre os mais variados temas como política, esporte, economia, dentre outras.

Todas essas informações não são exploradas em todo o seu potencial devido à capacidade humana de processamento manual ser limitada. Surge assim, a necessidade da criação de sistemas computacionais que sejam capazes de analisar

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

automaticamente o enorme volume de informações disponíveis na Web. Contudo, a maioria das informações na Web está representada em formato textual, escrita em linguagem natural, sendo destinada à consulta, análise e interpretação realizadas por pessoas.

O armazenamento em formato textual não é a forma mais apropriada para o processamento computacional, uma vez que não é estruturado e não expressa explicitamente os aspectos semânticos de seu conteúdo. Para que o processamento automático dessas informações seja realizado de forma mais eficaz, é necessário que essas informações sejam armazenadas em um formato estruturado permitindo que sejam interpretadas de maneira não ambígua. Com isso, é possível que tanto pessoas, quanto agentes computacionais possam analisar e extrair conhecimento útil. A transformação de construções sintáticas em semânticas é o principal objetivo da construção de bases de conhecimento.

A tarefa de construir Bases de Conhecimento é o processo de povoar um repositório de conhecimento com novos fatos extraídos a partir de uma ou mais fontes de informação. Esse processo requer o uso de técnicas de Extração de Informação (EI) e Processamento de Linguagem Natural (PLN) para analisar e transformar fontes de dados desestruturados (textos) em um formato estruturado. Essa tarefa apresenta diversos desafios e tem demandado interesse da comunidade de Inteligência Artificial ao longo dos anos, tendo sido propostos diversos trabalhos como o PANKOW [4], KnowItAll [6], Never-Ending Language Learning (NELL) [3], UMO-POW [14], OntoLP [17], Poronto [18], PSAPO [1], entre outros.

Para a realização do processo de construção de uma Base de Conhecimento é necessária a existência de uma estrutura básica que possa representar conceitos, relações e propriedades de um ou mais domínios. Esse alicerce pode ser representado utilizando uma ou mais Ontologia(s). A base de conhecimento então instancia os elementos presentes na ontologia para um determinado domínio.

O termo *Ontologia* surgiu originalmente na Filosofia, como uma área que trata da natureza e organização dos seres [8]. No contexto da Ciência da Computação, ontologias podem ser definidas como especificações explícitas e formais de uma conceitualização compartilhada [13]. Uma ontologia representa um domínio através de seus conceitos (classes), propriedades, relações, axiomas, hierarquia de conceitos (taxonomia de conceitos) e hierarquia de relações (taxonomia de relações).

Uma das principais tarefas para a construção e manutenção de Bases de Conhecimento é o *Povoamento de Ontologias* [9]. Povoamento de Ontologias é o processo de inserção de novas instâncias de classes, propriedades e/ou relações em uma ontologia existente [12]. Além disso, essa tarefa permite relacionar o conhecimento descrito em linguagem natural com ontologias, auxiliando o processo de geração de conteúdo semântico [16]. Por fim, a ontologia povoada pode ser usada em diversas aplicações, como gerenciamento de conteúdo, recuperação de informação, raciocínio automático, dentre outras.

Atualmente, ainda existem poucos trabalhos que investiram na criação de abordagens automáticas capazes de realizarem a tarefa de povoamento de ontologias, a partir de textos escritos em português. Em sua grande maioria os trabalhos adotam a língua inglesa, principalmente devido à quantidade de informações disponíveis nesse idioma e pela

quantidade de ferramentas que podem auxiliar no processo de povoamento.

Tentando suprir tal problema, o objetivo deste trabalho é propor e avaliar um método não supervisionado para o povoamento de ontologias utilizando a Web como grande corpus de trabalho. A abordagem proposta é capaz de extrair instâncias de classes ontológicas a partir de fontes textuais escritos em linguagem natural disponíveis na Web para o idioma português. O método proposto é guiado por uma ontologia de entrada que define quais conceitos devem ser povoados, e por um conjunto de padrões linguísticos usados para extrair e classificar um conjunto de termos candidatos a instâncias.

O restante deste artigo está organizado da seguinte forma: Na Seção 2, são apresentados os principais trabalhos relacionados com a tarefa de Povoamento de Ontologias, focando principalmente nos trabalhos que utilizam textos escritos em português como fonte de informações. Na Seção 3, são expostas as principais tecnologias e o método envolvido no seu desenvolvimento. Na Seção 4, os experimentos, resultados e discussões são ressaltados. Por fim, na Seção 5 são delineadas as conclusões e trabalhos futuros.

2. TRABALHOS RELACIONADOS

O processo de povoamento de ontologias de forma automática ou semiautomática depende diretamente do processo de aquisição do conhecimento. Várias propostas foram desenvolvidas, cada uma utilizando técnicas e métodos diferentes, que permitem encontrar informações contidas em documentos e armazená-las em diversas formas, como por exemplo, em ontologias.

Em [11] um processo para extrair informações estruturadas a partir de páginas da web é proposto. Seu objetivo é preencher uma ontologia de domínio de conhecimentos específicos que contenham afirmações declarativas em linguagem natural. Esse processo permite capturar informações semânticas contidas nos dicionários históricos textuais, ou seja, capturar entidades, relações e eventos e torná-los explícitos e disponíveis em uma ontologia.

Xavier e Lima [17] apresentam um estudo sobre a extração de uma estrutura ontológica contendo relações de hiponímia (é uma) e localização a partir da Wikipédia em língua portuguesa. A abordagem visa capturar a estrutura de categorias de enciclopédias, que contém um rico conteúdo semântico. As autoras fizeram um estudo de caso voltado para o domínio de Turismo, e a proposta objetiva o mapeamento da estrutura taxonômica da ontologia e as relações de localização entre as instâncias, além da extração de instâncias.

O trabalho apresentado por Drumond e Girardi [5] extrai estruturas taxonômicas a partir de textos, usando uma abordagem estatística, denominada Probabilistic Relational Hierarchy Extraction (PREHE), que faz a extração das estruturas através de reconhecimento de relações e outras técnicas de PLN e também valida seu estudo no domínio de Turismo.

Baségio [2] propõe uma abordagem para aquisição de estruturas ontológicas a partir de textos na língua portuguesa, mais especificamente, são extraídos conceitos e relações taxonômicas que servem como ponto de partida para o engenheiro de ontologia. O autor, para a validação de sua proposta, conduz experimentos sobre o domínio de Turismo.

Tomaz et. al. [14] propõem um método não supervisionado para o povoamento de ontologias a partir de textos

escritos em inglês disponíveis na Web. O método extrai termos candidatos a instância utilizando a Web como corpus e posteriormente combina diferentes medidas estatísticas e semânticas para classificar os termos extraídos. Os experimentos realizados obtiveram bons resultados utilizando uma ontologia de topo com classes de diferentes domínios.

A abordagem proposta neste trabalho baseia-se no trabalho de Tomaz et. al. [14] que foi desenvolvido originalmente para o idioma inglês. No melhor do conhecimento dos autores nenhum trabalho anterior investigou a aplicação de um processo de povoamento de ontologias similar ao proposto por Tomaz et. al.[14] para textos escritos em português. Essa lacuna motivou o desenvolvimento deste trabalho, no qual o método proposto se diferencia de outras propostas que lidam com textos em português por: (1) Utilizar a Web como grande corpus de trabalho para extração de termos candidatos a instâncias e (2) Avaliar diferentes medidas não supervisionadas para classificação de candidatos a instâncias.

3. ABORDAGEM PARA O POVOAMENTO DE ONTOLOGIAS PROPOSTA

O objetivo principal da abordagem proposta é extrair instâncias de classes ontológicas a partir de textos escritos em linguagem natural, em português, encontrados na Web. O processo de extração é guiado por uma ontologia de entrada que define os conceitos que devem ser povoados, e por um conjunto de padrões linguísticos independente de domínio adaptados de Hearst [7]. Esses padrões possuem dois papéis fundamentais: (1) Extratores, guiando o processo de extração de candidatos a instâncias; e (2) Discriminadores, sendo utilizados durante a classificação dos termos candidatos a instâncias. Os padrões de Hearst [7] foram originalmente utilizados para documentos em inglês. Neste trabalho, esses padrões foram traduzidos para o português.

A abordagem proposta é composta por 4 etapas principais: Coleta, Extração, Classificação e Povoamento. Uma visão geral das etapas e do fluxo de execução são apresentados na Figura 1 e detalhados nas seções a seguir.

3.1 Etapa 1 - Coleta dos Documentos

A primeira tarefa da abordagem proposta é selecionar uma classe c na ontologia de entrada. Após isso, utilizando um conjunto de padrões linguísticos independentes de domínio adaptados de Hearst [7], consultas são formuladas e aplicadas a um mecanismo de busca na Web para recuperação de um conjunto de documentos relevantes. São apresentados na Tabela 1, na 1ª coluna, os sete padrões linguísticos utilizados; e na 2ª coluna, exemplos de consultas formuladas para a classe Cidade.

Os padrões linguísticos listados na Tabela 1, em geral, são precedidos ou seguidos de exemplos de instâncias para a classe selecionada. O processo de formulação das consultas apresentadas na 2ª coluna é realizado da seguinte maneira: (1) o elemento *Classe* é substituído pelo rótulo da classe selecionada no singular, enquanto que o elemento *Classe(s)* é trocado pelo plural do rótulo da classe; (2) os elementos *Candidato* e *Candidatos* são removidos; e (3) o elemento ART é substituído pelos artigos indefinidos um ou uma de acordo com as regras gramaticais da língua portuguesa. Um detalhe importante é a presença de aspas nas consultas formuladas, indicando que a busca deve ser exata, ou seja, os documentos

Tabela 1: Padrões linguísticos com a respectiva consulta formulada

	Padrões Linguísticos	Consultas
1	<i>Classe(s)</i> tais como <i>Candidatos</i>	"cidades tais como"
2	tais <i>Classe(s)</i> como <i>Candidatos</i>	"tais cidades como"
3	<i>Candidatos</i> ou outro(a) <i>Classe(s)</i>	"ou outras cidades"
4	<i>Candidatos</i> e outro(a) <i>Classe(s)</i>	e "outras cidades"
5	<i>Classe(s)</i> incluindo <i>Candidatos</i>	"cidades incluindo"
6	<i>Classe(s)</i> especialmente <i>Candidatos</i>	"cidades especialmente"
7	<i>Candidato</i> é ART <i>Classe</i>	"é uma cidade"

só devem ser recuperados se possuírem exatamente a consulta usada.

Após a formulação das consultas, essas são aplicadas a um mecanismo de buscas na Web para recuperação de n documentos relevantes para cada consulta apresentada na Tabela 1. Neste trabalho, o Microsoft Bing ¹ foi utilizado como mecanismo de busca.

Após a etapa, um conjunto de documentos relevantes para a classe selecionada na ontologia de entrada está disponível para realização do processo de extração de candidatos a instâncias. Este processo é descrito na próxima seção.

3.2 Etapa 2 - Extração dos Candidatos a Instâncias

Após a etapa de coleta, cada um dos n documentos recuperados é processado. Para isso, a ferramenta CoGrOO ², foi utilizada para realização das tarefas de tokenização, divisão de sentenças, etiquetagem das classes gramaticais, stemming e identificação de sintagmas nominais.

O objetivo é extrair sentenças relevantes, ou seja, são extraídas apenas sentenças que possuem o padrão linguístico que originou a consulta, no qual *Candidatos* representa um conjunto de sintagmas nominais extraídos como candidatos a instâncias. Por exemplo, na sentença *Cidades tais como Nova Iorque, Tóquio, Londres, Paris e Hong Kong são grandes pólos financeiros*. os sintagmas nominais *Nova Iorque, Tóquio, Paris, Hong Kong* e *grandes pólos financeiros* são extraídos como candidatos a instâncias para a classe Cidade. Para o caso do padrão *Candidato é ART Classe*, o elemento Candidato denota um único sintagma nominal. Cada candidato a instância extraído, mantém a lista de padrões linguísticos responsáveis por sua extração sem repetição. Essa informação será usada posteriormente na fase de classificação dos candidatos a instâncias.

Usando sintagmas nominais é possível realizar a extração de palavras simples e compostas, aumentando assim, a cobertura dos candidatos a instâncias extraídos. Outros trabalhos como Etzioni et al. [6]; McDowell e Cafarella [10]; Tomaz et al.[14] também utilizaram com sucesso sintagmas nominais como candidatos a instâncias na tarefa de Povoamento de Ontologias.

¹<http://www.bing.com/>

²<http://ccsl.ime.usp.br/redmine/projects/cogroo>

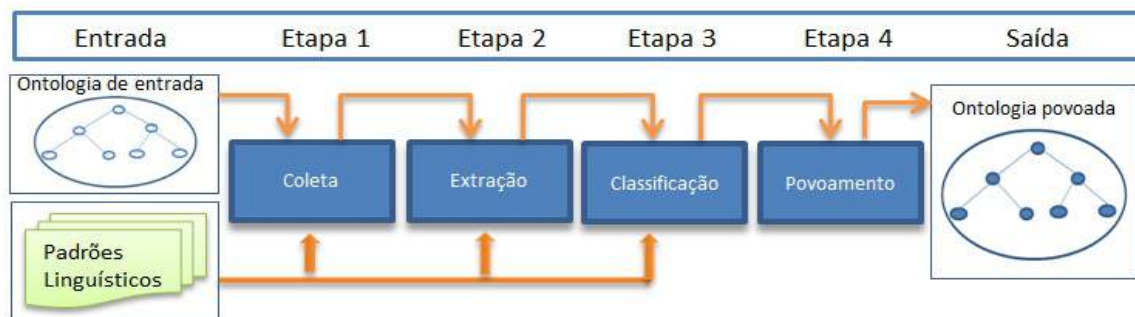


Figura 1: Visão geral da abordagem proposta

Com o objetivo de eliminar candidatos a instâncias inválidos ou repetidos, foram utilizadas alguns filtros:

a) *Filtragem Candidatos a Instâncias repetidos*: essa filtragem tem o objetivo de eliminar candidatos a instâncias que representam informações já existentes na ontologia de entrada. Por exemplo, dada a classe *Cidade* e um possível candidato a instância *Cidade*, ou variações, por exemplo, *ciudades*, *a cidade*, *uma cidade*, *na cidade*, entre outros. Esses candidatos são eliminados, pois, representam a própria classe selecionada. Candidatos a instâncias que sejam instâncias já presentes na classe em povoamento ou que sejam instâncias de classes disjuntas também são eliminados. Para aumentar a cobertura desse filtro, o algoritmo de *stemming* é utilizado para encontrar o radical das palavras.

b) *Filtragem de Candidatos a Instância sem valor semântico*: o objetivo deste filtro é eliminar candidatos a instâncias que não possuem valor semântico. No método proposto, candidatos que não possuem substantivo são removidos. Por exemplo, candidatos a instâncias como: *aqueles*, *a seguir*, *o quê* e *embora*, entre outros, são eliminados nessa filtragem.

c) *Filtragem Sintática*: a lista de candidatos a instâncias produzida por essa etapa não deve possuir candidatos repetidos. Caso um candidato seja extraído mais de uma vez, esse é inserido uma única vez na lista de candidatos a instâncias, sendo atualizada apenas a lista de padrões linguísticos distintos que o extraiu. Para melhorar o processo de identificação de redundâncias, aplica-se o algoritmo de *stemming* para evitar variações do singular e do plural. Além disso, candidatos que se diferem apenas pela presença de artigos, preposições e pronomes antes do primeiro substantivo, são mapeados para apenas uma única forma. Por exemplo, os candidatos a instâncias *O cavalo*, *um cavalo*, *seu cavalo*, *cavalo* e *cavalos*, são identificados como um único candidato a instância. A decisão de qual representação deve permanecer é realizada mensurando o grau de coocorrência entre cada candidato a instância e a classe selecionada.

Para mensurar a coocorrência, a medida Pontuação de Informação Mútua (PMI) apresentada na Equação 1 e os padrões linguísticos listados na Tabela 1 são utilizados. Maiores detalhes sobre a medida de PMI são apresentados na próxima seção. O candidato *c* com maior valor de coocorrência é inserido na lista de candidatos a instâncias, além disso, a lista de padrões linguísticos dos candidatos removidos são adicionadas na lista de padrões linguístico de *c*.

Ao final desta etapa, após o processamento de todos os documentos coletados, uma lista formada pelos candidatos a instâncias extraídos para a classe selecionada e os padrões

Tabela 2: Lista de candidatos a instância com padrões linguísticos que o extraíram

Candidato a instância	Padrões linguísticos que o extraíram
Alexandria	<i>Candidatos</i> ou outro(s) <i>Classe(s)</i> <i>Candidato</i> é ART <i>Classe</i> <i>Classe(s)</i> incluindo <i>Candidatos</i>
Teresópolis	<i>Classe(s)</i> tais como <i>Candidatos</i> <i>Classe(s)</i> incluindo <i>Candidatos</i>
Viçosa	<i>Candidato</i> é ART <i>Classe</i>
Belo Horizonte	<i>Candidatos</i> ou outro(s) <i>Classe(s)</i> <i>Classe(s)</i> tais como <i>Candidatos</i> <i>Candidato</i> é ART <i>Classe</i> <i>Candidatos</i> e outro(s) <i>Classe(s)</i>

linguísticos que o extraíram, é produzida. É ilustrado na Tabela 2 exemplos de candidatos a instâncias extraídos para a classe *Cidade* e os respectivos padrões que o extraíram.

3.3 Etapa 3 - Classificação dos Candidatos a Instâncias

Após a extração dos candidatos a instâncias, é necessário avaliar a confiabilidade de cada candidato classificado. Esta etapa tem por objetivo avaliar cada candidato identificado pela etapa Extração dos Candidatos a Instâncias e atribuir um grau de confiança a cada candidato a instância extraído.

Para isso, neste trabalho foram avaliadas três variações tradicionais da medida de Pontuação de Informação Mútua, do inglês *Pointwise Mutual Information* (PMI). O PMI é uma medida estatística cujo objetivo é mensurar o grau de coocorrência entre dois termos. Em geral, as medidas estatísticas sofrem com o problema de esparsidade dos dados [4], ou seja, dependendo da fonte de informação utilizada, os dados disponíveis nem sempre apresentam um indicativo de sua relevância, refletindo assim em uma baixa performance, principalmente quando utiliza-se palavras relativamente raras.

Para resolver tal problema, diversos autores [4], [6], [10], [14] demonstraram que a utilização de medidas estatísticas explorando a grande quantidade de dados disponíveis na Web apresenta-se como uma solução viável.

Diante disso, neste trabalho, a medida de PMI objetiva mensurar o grau de coocorrência entre uma classe (*c*) e cada um dos seus candidatos a instâncias (*ci*) usando um conjunto de padrões linguísticos (*P*). Para isso, consultas são formuladas utilizando *c*, *ci* e cada padrão linguístico $p \in P$.

Tabela 3: Lista de candidatos a instância com padrões linguísticos que o extraíram

(1) Padrões Linguísticos	(2) Consultas
<i>Classe(s)</i> tais como <i>Candidatos</i>	cidades tais como Salvador
tais <i>Classe(s)</i> como <i>Candidatos</i>	tais cidades como Salvador
<i>Candidatos</i> ou outro(s) <i>Classe(s)</i>	Salvador ou outras cidades
<i>Candidatos</i> e outro(s) <i>Classe(s)</i>	Salvador e outras cidades
<i>Classe(s)</i> incluindo <i>Candidatos</i>	cidades incluindo Salvador
<i>Classe(s)</i> especialmente <i>Candidatos</i>	cidades especialmente Salvador
<i>Candidato</i> é ART <i>Classe</i>	Salvador é uma cidade

Posteriormente essas consultas são aplicadas a algum mecanismo de busca na Web com o objetivo de obter a quantidade de ocorrências de cada consulta executada. São ilustradas na Tabela 3 as consultas formuladas para o cálculo do PMI utilizando a classe Cidade, o candidato a instância Salvador e os padrões linguísticos apresentados na Tabela 1.

O processo de formulação das consultas apresentadas na Tabela 3 é realizado de maneira semelhante à etapa de Extração de Candidatos. A diferença está no uso do elemento Candidato(s) que é substituído pelo candidato a instância em avaliação.

Diversos trabalhos utilizam diferentes variações de fórmulas aplicadas no cálculo do PMI [15], [4], [6], [10], [14]. Em particular, McDowell e Cafarella [10] e Tomaz et al. [14], apresentam um estudo realizado para comparar diferentes variações para o cálculo do PMI aplicados no Povoamento de Ontologias na Web para textos escritos em inglês. Seguindo a mesma ideia, um experimento foi realizado com o objetivo de avaliar as 3 variações do PMI previamente investigadas por [10] e [14].

Nas equações 1, 2 e 3 são apresentadas as três variações da medida de PMI analisadas neste trabalho. Nestas equações, $hits(ci, c, p)$ representa o número de ocorrências retornadas pelo mecanismo de busca na Web para a consulta formada por um candidato a instância ci , uma classe c , usando o padrão linguístico p . Enquanto isso, $hits(ci)$ e $hits(c)$ representam o total de ocorrência do candidato a instância ci e da classe c isoladamente.

PMI Strength: Essa variação do PMI é calculada pelo somatório de todas as coocorrência retornadas pela consulta $hits(c, ci, p)$. Para isso, consultas são formuladas para o par (c, ci) e cada padrão linguístico $p \in \mathcal{P}$ listados na Tabela 1. Na Equação 1 é apresentado como é realizado o cálculo desta variação de PMI.

$$PMI_{Strength} = \sum_{p \in \mathcal{P}} hits(c, ci, p) \quad (1)$$

PMI Str-INorm-Thresh: Segundo McDowell e Cafarella [10], a variação *PMI Strength* pode ser tendenciosa para instâncias muito frequentes. Outras instâncias para a classe selecionada podem não ser tão frequentes. Tentando solucionar esse problema, algumas variações utilizam como fator de normalização o total de ocorrências retornado pelo

$hits(ci)$. Contudo, mesmo usando esse fator de normalização em casos onde o candidato à instância é raro, o problema da tendenciosidade persiste. Diante disso, o fator de normalização precisa ter um valor mínimo definido. Nesta versão, o fator de normalização utilizado é resultado do maior valor entre $hits(ci)$ do candidato a instância em classificação e o 25º percentil da distribuição de $hits(ci)$ de todos os candidatos a instância para a classe c selecionada. Na Equação 2 é apresentado como é realizado o cálculo desta variação de PMI.

$$PMI_{Str-INorm} = \frac{\sum_{p \in \mathcal{P}} hits(c, ci, p)}{\max(hits(ci), Percen_{25})} \quad (2)$$

PMI Str-ICNorm-Thresh: Seguindo a mesma ideia de normalização do *PMI Str-INorm-Thresh*, esta variação utiliza, como fator de normalização, ambos, o candidato a instância ci e a classe c . Na Equação 3 é apresentada como é calculada esta variação.

$$PMI_{Str-ICNorm} = \frac{\sum_{p \in \mathcal{P}} hits(c, ci, p)}{\max(hits(ci), Percen_{25}) * hits(c)} \quad (3)$$

Além das 3 variações de PMI apresentadas, neste trabalho, a heurística de Número de Padrões que Extraíram (NPE) proposta por [14] foi analisada. A heurística de NPE é baseada na hipótese de quanto mais padrões linguísticos distintos forem responsáveis pela extração de um candidato a instância, mais forte é a evidência de que esse candidato seja realmente uma instância válida para a classe em povoamento. Baseado nessa hipótese, a heurística de NPE mensura quantos padrões linguísticos distintos extraíram o candidato à instância avaliado. O intervalo de valores que essa heurística pode assumir varia de 1 ao total de padrões linguísticos utilizados pelo método para a classe selecionada. Por exemplo, na Tabela 2, o candidato a instância Alexandria possui $NPE = 3$, já que foi extraído por 3 padrões linguísticos, enquanto que o candidato a instância Viçosa possui $NPE = 1$.

Ao final desta etapa todos os candidatos a instâncias possuem um valor de confiança atribuído por cada uma das 3 variações de PMI e também pela heurística de NPE. Na seção 4 são apresentados e discutidos os experimentos realizados para investigar qual das 4 medidas de classificação apresenta a melhor performance na tarefa de classificar os candidatos a instâncias.

3.4 Etapa 4 - Povoamento

Nesta última etapa, o objetivo é decidir quais desses candidatos a instâncias serão promovidas a instâncias da classe selecionada, sendo assim utilizadas para povoar a ontologia de entrada. Diante disso, a escolha de um limiar é muito importante principalmente levando em consideração que essa escolha impacta diretamente na taxa de acerto e na cobertura do método.

Em geral, a escolha do limiar é realizada empiricamente, sendo promovida a instância apenas candidatos que possuem um valor maior do que o limiar escolhido. Conforme apontado por Tomaz et al. [14], a medida de PMI possui valores variando em ordem de grandeza diferentes dependendo da classe selecionada, do candidato a instância em avaliação e dos padrões linguísticos utilizados. Diante disso, estimar um valor para o limiar tornou-se inviável. Ao invés disso, assim

como Tomaz et al.[14] optou-se por promover a cada iteração os n melhores candidatos a instâncias (Top n) ordenados com base nos valores de confiança de cada uma das medidas de classificação descritas na seção anterior.

4. EXPERIMENTO E DISCUSSÕES

Para demonstrar a eficácia da abordagem proposta, um experimento comparando as 3 variações da medida de PMI e também a heurística de NPE foi realizado. Esse experimento vislumbra avaliar cada uma das etapas do método proposto, focando principalmente na análise comparativa entre as 4 medidas apresentadas aplicadas na tarefa de classificação dos candidatos a instâncias das classes ontológicas.

4.1 Configurações do Experimento

O corpus utilizado neste experimento é formado por um conjunto de *snippets* coletados utilizando o mecanismo de busca Bing. *Snippets* são textos simples que em geral possuem as palavras chave que formam a consulta aplicada, apresentando uma prévia da informação contida nos documentos recuperados. Essas prévias, mesmo possuindo um tamanho reduzido, são informativos o suficiente para extrair conhecimento relacionado com a consulta aplicada sem a necessidade de processar o documento inteiro [14].

Para este experimento 15 classes foram selecionadas em uma ontologia de topo customizada: *Cidade*, *País*, *Pássaro*, *Peixe*, *Sintoma*, *Esporte*, *Inseto*, *Mamífero*, *Doença*, *Universidade*, *Ator*, *Atriz*, *Filme*, *Rio* e *Hotel*.

A coleta de documentos foi efetuada de forma automática e buscava 1000 documentos/dia. No período de 19/09/2014 a 20/11/2014 foram recuperados para as 15 classes selecionadas um total de 62.409 documentos relevantes.

A medida utilizada para avaliar os experimentos realizados foi a medida de *Precisão*. No contexto deste trabalho, a precisão pode ser definida pela razão entre a quantidade de instâncias corretas extraídas pelo total de instâncias recuperadas. A Equação 4 foi usada para calcular a precisão dos experimentos.

$$Precisao(Top_N) = \frac{total_de_instancias_corretas}{N} \quad (4)$$

Nesta equação, N é a quantidade de termos candidatos promovidos a instância.

A avaliação da precisão neste experimento foi realizada variando diferentes limiares. O limiar definiu o valor dos N candidatos a instâncias (Top N) ordenados com base na medida de classificação utilizada. Neste experimento foram utilizados quatro limiares: Top 10, Top 50, Top 100 e Top 200. Os limiares Top 10 e Top 50 são mais restritivos, ou seja, poucas instâncias são extraídas. Já os limiares Top 100 e Top 200 visam promover um número maior de instâncias por iteração.

Para o cálculo da Precisão é necessário analisar cada termo candidato extraído e verificar se ela é realmente uma instância para a classe ao qual foi atribuída. Para isso, 27 humanos foram responsáveis pelo processo de validação de cada um dos Top N melhores candidatos a instâncias classificados.

Para promover uma comparação justa entre as medidas de classificação avaliadas, as etapas de coleta e extração de candidatos a instâncias foram executadas uma única vez. Em seguida, os conjuntos de candidatos a instâncias extraídos para cada uma das 15 classes selecionadas foram passados

como parâmetros de entrada para a etapa de classificação que utilizou as 3 medidas de PMI e a heurística de NPE individualmente para classificar os candidatos a instâncias.

4.2 Resultados e Discussões

Os gráficos apresentados nas Figuras 2, 3, 4 e 5 demonstram os resultados obtidos para as 15 classes selecionadas na ontologia de entrada. Em todos os gráficos são realizadas as comparações entre as medidas de PMI Strength, PMI Str-INorm-Thresh, PMI Str-ICNorm-Thresh e a heurística de NPE.

Analisando os resultados apresentados nas Figuras 2, 3, 4 e 5 é possível observar que a abordagem proposta é capaz de extrair uma grande quantidade de instâncias corretas para a maioria das classes selecionadas. Nos limiares mais restritivos Top 10 e Top 50 maiores valores de precisão foram alcançados, enquanto que nos limiares mais abrangentes Top 100 e Top 200 houve uma tendência natural de perda de precisão. Observa-se também que existe uma variação de precisão entre as classes selecionadas, isso ocorre devido a fatores como complexidade do domínio e coocorrência de instâncias da classe selecionada com os padrões linguísticos utilizados. Por exemplo, nas classes Cidade e País que são de domínio geral, as medidas avaliadas obtiveram uma alta taxa de precisão em relação a outras classes de domínio mais específico como Inseto, Passáro e Sintoma.

Um fator importante a ser ressaltado é a falta de confiabilidade das informações presentes na Web. A maioria dessas informações é escrita por pessoas que em geral não são especialistas do domínio abordado. Quanto mais complexo o domínio analisado, maior a probabilidade de encontrar informações erradas.

Considerando o limiar Top 10, como apresentado na Figura 2, a heurística de NPE apresentou uma maior taxa de precisão em 9 das 15 classes analisadas obtendo uma média geral de 67% de precisão. As variações PMI Str-INorm-Thresh e PMI Str-ICNorm-Thresh apresentaram maiores taxas de precisão em 8 das 15 classes, ficando ambas com uma média de 60% de precisão. Por fim, a variação PMI Strength apresentou melhores taxas de precisão apenas em 3 classes, ficando com uma média de precisão geral de 49%.

Analisando os limiares Top 50 e Top 100, como apresentado nas Figuras 3 e 4, o mesmo comportamento foi observado. A heurística de NPE continuou melhor com uma média geral de precisão de 60% no Top 50 e 48% no Top 100. Em segundo lugar as variações PMI Str-INorm-Thresh e PMI Str-ICNorm-Thresh continuaram empatadas com 55% de precisão média no Top 50 e 41% no Top 100. Por último, a variação de PMI Strength obteve 52% de precisão média no Top 50 e 40% no Top 100.

No limiar Top 200, apresentado na Figura 5, as 4 medidas ficaram com médias de precisão muito próximas com uma diferença de apenas 1% para a heurística de NPE que obteve precisão média de 49%. Enquanto isso, as 3 variações de PMI ficaram empatadas com uma média de 48% de precisão.

Com base nos experimentos executados, conclui-se que a heurística de NPE obteve melhores resultados do que as 3 variações da medida de PMI em todos os limiares. Contudo, a diferença entre elas foi diminuindo à medida que o limiar foi aumentando. As variações PMI Str-INorm-Thresh e PMI Str-ICNorm-Thresh obtiveram melhores resultados do que a variação PMI Strength. Tal resultado indica que os fatores de normalização usados pelo PMI Str-INorm-Thresh e PMI

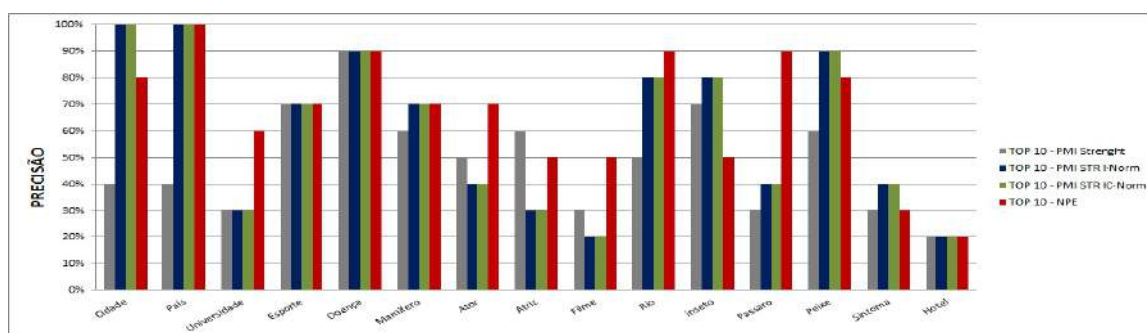


Figura 2: Resultados do Experimento no limiar Top 10.

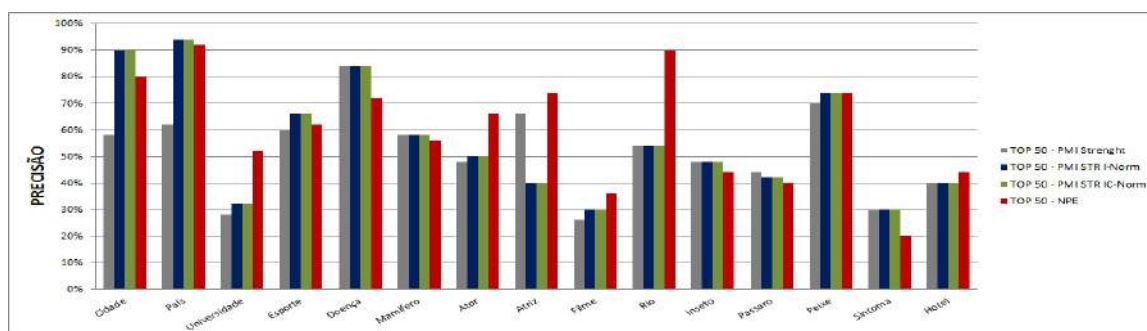


Figura 3: Resultados do Experimento no limiar Top 50.

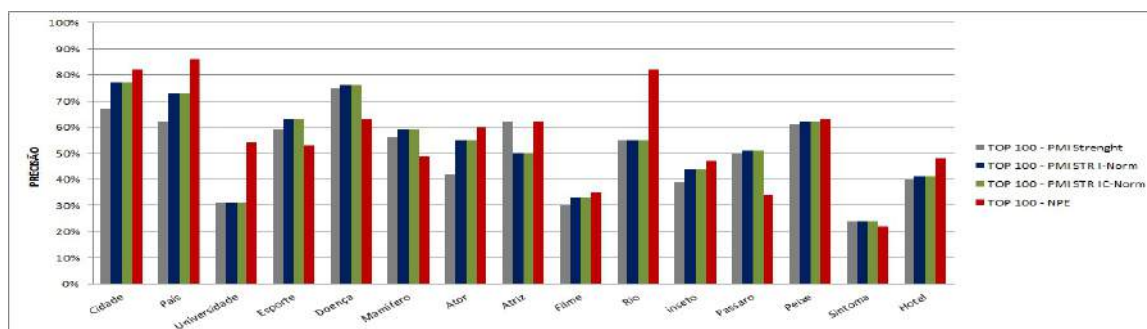


Figura 4: Resultados do Experimento no limiar Top 100.

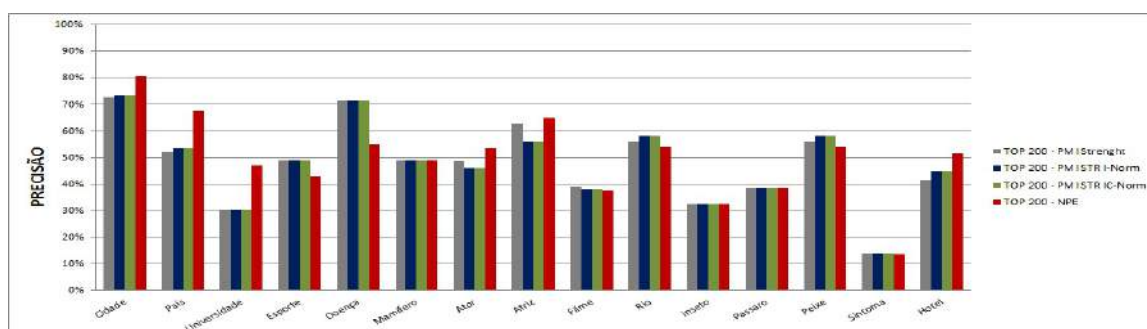


Figura 5: Resultados do Experimento no limiar Top 200.

Str-ICNorm-Threshold apresentaram um impacto positivo nos resultados.

5. CONCLUSÕES

Este trabalho apresentou uma abordagem não supervisionada para o povoamento de ontologias a partir de textos escritos em linguagem natural em português utilizando a Web como grande fonte de informações. A relevância desse problema é decorrente da abundância de informações desestruturadas presentes na Web e da aquisição de conhecimento estruturado a partir de textos desestruturados para criação de bases de conhecimento. Além disso, acrescenta-se os poucos estudos que investigam a exploração de textos escritos em português como fonte de informação.

Após a análise dos resultados é possível concluir que a abordagem para o povoamento de ontologias na Web foi capaz de extrair uma grande quantidade de instâncias corretas para a maioria das quinze classes selecionadas. Esses resultados são encorajadores para que a abordagem proposta possa ser evoluída para atingir melhores resultados e também expandida para contemplar outros aspectos na ontologia de entrada como relações entre classes e propriedades.

Como trabalhos futuros sugere-se: (1) Propor uma medida combinada que integre diferentes medidas e heurísticas para a classificação dos candidatos a instâncias extraídas; (2) Integrar um módulo para identificação de sinônimos para a filtragem de candidatos a instâncias que possuem o mesmo valor semântico; e (3) Explorar recursos semânticos disponíveis para o idioma português, por exemplo, o OpenWordnet-PT³.

6. REFERÊNCIAS

- [1] C. G. d. F. Alves. Um Processo Independente de Domínio para o Povoamento Automático de Ontologias a partir de Fontes Textuais. 2013.
- [2] T. L. Baségio. Uma Abordagem Semi-automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil. pages 1–124, 2007.
- [3] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *In AAAI*, 2010.
- [4] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. *Proceedings of the 13th conference on World Wide Web - WWW '04*, page 462, 2004.
- [5] L. Drumond and R. Girardi. Extracting ontology concept hierarchies from text using markov logic. In *Proceedings of the 2010 ACM Symposium on Applied Computing, SAC '10*, pages 1354–1358, New York, NY, USA, 2010. ACM.
- [6] O. Etzioni, S. Kok, S. Soderland, M. Cafarella, A. m. Popescu, D. S. Weld, D. Downey, T. Shaked, and A. Yates. Web-Scale Information Extraction in KnowItAll (Preliminary Results). pages 100–110, 2004.
- [7] M. A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. pages 23–28, 1992.
- [8] A. Maedche and S. Staab. Ontology learning for the semantic web. *Intelligent Systems, IEEE*, 16(2):72–79, Mar 2001.
- [9] D. Maynard, Y. Li, and W. Peters. Nlp techniques for term extraction and ontology population. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*, pages 107–127, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.
- [10] L. K. McDowell and M. Cafarella. Ontology-driven, unsupervised instance population. *Web Semant.*, 6(3):218–236, Sept. 2008.
- [11] E. N. Motta. Preenchimento Semi-automático de Ontologias de Domínio a Partir de Textos em Língua Portuguesa. Master's thesis.
- [12] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos. Ontology population and enrichment: State of the art. In G. Paliouras, C. Spyropoulos, and G. Tsatsaronis, editors, *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, volume 6050 of *Lecture Notes in Computer Science*, pages 134–166. Springer Berlin Heidelberg, 2011.
- [13] S. Studer, S. Rudi, H.-P. Schnurr, and Y. SURE. Knowledge processes and ontologies. *IEEE Intelligent Systems*, 1(16):26–34, 2001.
- [14] H. Tomaz, R. Lima, J. Emanuel, and F. Freitas. An unsupervised method for ontology population from the web. In J. Pavón, N. Duque-Méndez, and R. Fuentes-Fernández, editors, *Advances in Artificial Intelligence – IBERAMIA 2012*, volume 7637 of *Lecture Notes in Computer Science*, pages 41–50. Springer Berlin Heidelberg, 2012.
- [15] P. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. 2001.
- [16] D. C. Wimalasuriya. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323, Mar. 2010.
- [17] C. C. a. Xavier and V. L. S. d. Lima. A Semi-Automatic Method for Domain Ontology Extraction from Portuguese Language Wikipedia's Categories.
- [18] F. M. Zahra, D. R. Carvalho, and A. Malucelli. Poronto : ferramenta para construção semiautomática de ontologias em português Poronto : herramienta para construcción semiautomática de ontologías en portugués. *journal of Health Informatics*, 5(2):52–59, 2013.

³<http://logics.emap.fgv.br/wn/>

Projeto e implementação do framework Outer-tuning: auto sintonia e ontologia para bancos de dados relacionais

Alternative title: Design of the Outer-tuning framework: self-tuning and ontology for relational databases

Rafael Pereira de Oliveira
Pontifícia Universidade
Católica do Rio de Janeiro -
PUC-Rio
Rio de Janeiro, Brasil
rpoliveira@inf.puc-rio.br

Sérgio Lifschitz
Pontifícia Universidade
Católica do Rio de Janeiro -
PUC-Rio
Rio de Janeiro, Brasil
sergio@inf.puc-rio.br

Ana Carolina Almeida
Universidade do Estado do
Rio de Janeiro
Rio de Janeiro, Brasil
ana.almeida@ime.uerj.br

Edward Hermann
Haeusler
Pontifícia Universidade
Católica do Rio de Janeiro -
PUC-Rio
Rio de Janeiro, Brasil
hermann@inf.puc-rio.br

RESUMO

Neste trabalho discutimos a construção e o projeto de um *framework* - chamado Outer-Tuning - que dá apoio à sintonia fina (semi) automática de sistemas de bancos de dados por meio de uma ontologia específica. São apresentados os aspectos arquiteturais baseados em componentes, projeto de interface e, principalmente, a abordagem adotada para contemplar tanto uma ontologia de domínio como uma ontologia de tarefas em um sistema de informação integrador. Discutimos também aspectos relacionados com a máquina de regras de inferência e questões relacionadas ao uso de linguagens lógicas. Por fim, alguns resultados experimentais permitem uma avaliação preliminar das contribuições esperadas pelo *framework*.

Palavras-Chave

Framework, design, sintonia fina, banco de dados, ontologia

ABSTRACT

In this paper, we discuss the architectural project and construction of a framework called Outer-Tuning, which supports (semi) automatic tuning of database systems through a specific ontology. The architectural aspects component-based, interface design, and especially the approach taken to include both a domain ontology and an ontology tasks in

an integrative information system. We also discuss aspects of machine inference rules and issues related to the use of logical languages. Finally, some experimental results allow an assessment of the expected framework contributions.

Categories and Subject Descriptors

H [Information Systems]; H.2 [DATABASE MANAGEMENT]; H.2.7 [Database Administration]

General Terms

Design, experimentation

Keywords

Framework, design, tuning, databases, ontology

1. INTRODUÇÃO

O trabalho de administração e sintonia fina (*tuning*) de bancos de dados é complexo e exige especialização e conhecimentos fundamentais nesta área da computação. Vários sistemas foram desenvolvidos para dar apoio ao DBA (database administrator) nesta atividade, cujo objetivo principal é garantir disponibilidade e bom desempenho. Em particular, alguns sistemas permitem a realização de tarefas de sintonia fina de maneira (semi) automática [20][6]. Em sua maioria, são sistemas desenvolvidos a partir do conhecimento de um DBA especialista, que geram algoritmos e heurísticas de manutenção do sistema de banco de dados.

O principal objetivo da sintonia fina é buscar um melhor desempenho para o banco de dados. Para isso, realizam-se ajustes de suas configurações, parâmetros e projeto físico, seleção de estruturas de acesso, duplicação de estruturas físicas, sempre de acordo com a carga de trabalho executada no banco. Existem diversas propostas de ferramentas

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

que auxiliam a sintonia fina através da automatização de estratégias para a seleção de índices e visões materializadas, particionamento, reescrita de consultas, entre outras.

Apresenta-se aqui o resultado do projeto de arquitetura de *software*, desenvolvimento e avaliação do *framework* Outer-Tuning [1] ferramenta baseada em ontologia e criada para apoiar DBAs (e possivelmente outros usuários) nas escolhas envolvidas na atividade de sintonia fina. Em particular, discutimos os principais requisitos funcionais, o fluxo operacional e as diversas fases de execução necessárias para atender os requisitos. Também discutimos em detalhes os componentes propostos para a arquitetura, as alternativas de projeto e as escolhas efetuadas, incluindo a interface do usuário. Por estarmos utilizando uma ontologia especificamente construída para apoio à tarefa de sintonia fina em bancos de dados relacionais, discutimos também o uso de linguagens como DATALOG, OWL-DL e SWRL e sua integração na construção e implementação do *framework*.

Para a validação do *framework* foi realizada a instanciação bem como a extensão da ontologia de sintonia fina usada por ele, através da inclusão de duas novas heurísticas. A partir de uma carga de trabalho, essas novas heurísticas realizam a seleção e criação de visões materializadas. Com o uso de um *benchmark*, foi realizada uma avaliação experimental da qualidade dos resultados obtidos pelo Outer-Tuning, através da documentação, avaliação e comparação das ações de sintonia fina inferidas, como resultado de sua execução, por suas heurísticas.

A Seção 2 traz os fundamentos e conceitos envolvidos neste artigo, além de alguns trabalhos relacionados à nossa pesquisa. Logo após, a Seção 3 documenta o nosso método de pesquisa. A Seção 4 apresenta o projeto de arquitetura do *framework* Outer-Tuning e as principais escolhas realizadas durante a sua implementação. Discute-se o uso de uma ontologia para inferir ações de sintonia fina, a integração de uma máquina de regras com o *framework* e os principais componentes de *software* envolvidos. Em seguida, a Seção 5 traz uma avaliação preliminar do Outer-Tuning através da execução de duas heurísticas de sintonia fina automática envolvendo a seleção de visões materializadas. Por fim, a Seção 6 lista as principais conclusões e trabalhos futuros.

2. FUNDAMENTOS E TRABALHOS RELACIONADOS

O trabalho apresentado por [1] trouxe a proposta de um *framework*, chamado Outer-Tuning, para apoiar o trabalho de sintonia fina (semi)automática em sistemas de bancos de dados. A ideia básica da ferramenta Outer-Tuning consiste em oferecer mecanismos que permitam ao DBA uma tomada de decisão mais consistente com relação às possíveis ações de sintonia fina, baseada em alternativas e respectivos custos. Devido ao escopo da pesquisa e limitações de tempo não foi realizado o projeto por completo, consequentemente, não houve implementação do *framework*. Alguns testes preliminares foram realizados por meio de simulações utilizando o *software* Protégé¹.

A pesquisa apresentada nesse artigo tem como ênfase a descrição do projeto de arquitetura do *framework* Outer-Tuning. A partir da ontologia específica de sintonia fina em bancos de dados relacionais, proposta em [1] e estendida em

[17], buscamos atender os requisitos - listados na Seção 4.1 - e os desafios de especificação dos componentes e integração com uma máquina de regras. Apesar do *framework* Outer-Tuning ter sido definido usando como base ações de sintonia fina sobre índices, buscamos nesta pesquisa estender a discussão para o uso de visões materializadas (VM), de forma a não particularizar as soluções até então obtidas e permitir uma melhor compreensão das vantagens do uso de uma ontologia em todo o processo. Mais detalhes da ontologia de sintonia fina podem ser encontrados em [17].

O conceito de visão, em bancos de dados relacionais, diz respeito basicamente à uma consulta definida sobre relações do banco de dados. A definição da visão fica armazenada no catálogo e, enquanto um objeto do banco de dados, pode ser utilizado em consultas da mesma forma que as relações que pertencem ao esquema relacional. Já visões materializadas são consultas que têm não somente suas definições como também seus resultados fisicamente armazenados no banco de dados. As visões materializadas são implementadas pelos SGBDs mais populares como PostgreSQL, SQL Server, Oracle, entre outros [9].

Visões materializadas possuem custos. Não é necessário ter apenas espaço em disco suficiente, mas também requer o custo de criação, manutenção e execução de uma consulta SQL utilizando uma VM. Quando uma tabela é atualizada, consequentemente o conteúdo da visão materializada se torna desatualizado. Logo, surge a necessidade de seleção do conjunto de visões materializadas hipotéticas que tragam o maior benefício e os menores custos possíveis, dada uma carga de trabalho.

A formalização da escolha do conjunto de visões materializadas consideradas viáveis é dada pela seguinte definição: “Dado um esquema de um banco de dados R , armazenado em espaço D e uma carga de trabalho de consultas Q , escolher um conjunto de consultas V sobre R para ser materializado de forma que o somatório do tamanho dos elementos de V seja menor que D ” [9].

Além da questão de viabilidade, deve-se considerar quais VMs podem ser úteis para melhorar o desempenho de um sistema de banco de dados. A escolha do conjunto de consultas, cujo custo de execução da carga de trabalho é minimizado, dado uma restrição de espaço em disco e um custo de manutenção, é um problema com complexidade exponencial [19]. Logo, não é possível materializar todas as possíveis visões hipotéticas a melhorar o desempenho por pelo menos dois motivos: (i) o espaço em disco necessário para materializar todas as opções inviabiliza esta escolha e (ii) o custo de manter as visões materializadas atualizadas [9] é proibitivo.

Especificamente para a seleção automática e semiautomática de VMs, existem ferramentas que utilizam diferentes métodos de seleção do conjunto de consultas para a criação dessas estruturas de acesso. Entre elas, podemos citar: (i) pesquisas que utilizam heurísticas gulosas [24], [2], [8]; (ii) métodos de seleção baseados em heurísticas randômicas [13], [25], [21], [22]; (iii) abordagens baseadas em heurísticas genéticas e evolucionárias [24], [23], [12]; (iv) heurísticas híbridas [25], que são uma mistura de abordagens genética e randômica; e (v) heurísticas específicas [7].

Apesar de utilizarem VMs para realizar sintonia fina, nenhuma das abordagens encontradas se propõe a fornecer ao DBA uma justificativa das decisões e tornar explícito o raciocínio utilizado na atividade de sintonia fina. Também não possuem a possibilidade de comparação ou inclusão de novas

¹<http://protege.stanford.edu/> acesso em 18/02/2015.

heurísticas sem a necessidade de uma alteração do seu código fonte, tornando a comparação e adaptação das estratégias à carga de trabalho difícil e trabalhosa.

Todas estas características são oferecidas pelo *framework* Outer-Tuning. Trata-se de uma abordagem que faz uso de uma ontologia de aplicação que permite a inferência de ações de sintonia fina. A ontologia de domínio descreve os conceitos envolvidos na tarefa de sintonia fina e a ontologia de tarefas contém heurísticas capazes de analisar a carga de trabalho instanciada na ontologia de domínio. Neste artigo discutimos, exatamente, o projeto e arquitetura do *framework* Outer-Tuning, seus componentes e alternativas de arquitetura, com o detalhamento das decisões tomadas para efetiva implementação da ferramenta.

3. MÉTODO DE PESQUISA

Inicialmente, foi realizada uma pesquisa bibliográfica para: (i) investigar os passos necessários para o processo de sintonia fina; (ii) conhecer ferramentas de sintonia fina que realizem a seleção de visões materializadas; (iii) levantar possíveis arquiteturas para um *framework* de sintonia fina; (iv) levantar métodos de desenvolvimento de *frameworks*; e (v) avaliar máquinas de regras passíveis de integração com o *framework* e compatíveis com a ontologia de sintonia fina. Em um segundo momento, foi realizado o planejamento dos passos necessários para a inferência de ações de sintonia fina (Figura 1).

Para o desenvolvimento do projeto de arquitetura do *framework*, buscou-se utilizar a metodologia de projeto dirigido por *hot spot* (*Hot Spot Driven Design*) [18], tendo como base a ontologia de tarefa. Os *hot spots* são as partes flexíveis do *framework*. A essência dessa metodologia é identificar os *hot spots* na estrutura de classes de um domínio e, a partir disto, construir o *framework*.

Por fim, para a avaliação do *framework* foi realizada uma instanciação e extensão da ontologia de sintonia fina, onde foram incluídas duas novas heurísticas para, dada uma carga de trabalho, realizarem a seleção e criação de visões materializadas. Com a utilização de um *benchmark* para gerar a carga de trabalho, o Outer-Tuning foi então executado e teve as ações de sintonia fina inferidas pelas suas heurísticas (resultado de sua execução) documentadas, avaliadas e comparadas.

4. FRAMEWORK OUTER-TUNING

As fases de execução do *framework* foram definidas de acordo com os requisitos funcionais levantados a partir da análise dos requisitos do Outer-Tuning e pela avaliação das ferramentas de seleção de visões materializadas citadas na Seção 2. Após a definição das fases, foi desenvolvido o projeto de arquitetura do *framework*. Também é apresentado na Seção 4.4 uma análise sobre o uso de linguagens lógicas para realizar tarefas de sintonia fina.

4.1 Requisitos funcionais do framework Outer-Tuning

Durante a definição do escopo do *framework* Outer-Tuning, levantou-se requisitos desejáveis. Entre eles, que a partir da ontologia de sintonia fina proposta, fosse capaz de:

- (i) Capturar a carga de trabalho de um banco de dados

de forma contínua; (ii) Extrair da carga de trabalho os conceitos necessários à ontologia de domínio; (iii) Instanciar os conceitos extraídos da carga de trabalho na ontologia; (iv) Dar a oportunidade ao DBA (ou outro usuário) de escolher qual(is) heurística(s) de sintonia fina deseja executar durante o processo de sintonia fina; (v) Executar as heurísticas escolhidas pelo DBA e inferir as ações de sintonia fina previstas nas heurísticas; (vi) Exibir ao DBA justificativas semânticas sobre as ações inferidas pelas heurísticas escolhidas; (vii) Executar no banco de dados as ações de sintonia fina inferidas pelas heurísticas, de forma automática ou supervisionada pelo DBA. A partir dos requisitos listados, foram definidas as fases de execução do Outer-Tuning.

4.2 Fases de execução do framework Outer-Tuning

O fluxo definido para o Outer-Tuning (Figura 1) se inicia com o monitoramento do banco de dados de forma não intrusiva e contínua (2), que captura a carga de trabalho (3) submetida ao banco de dados (1) por seus usuários. Em seguida são extraídos (4) da carga de trabalho os conceitos, cujos valores são necessários para que as heurísticas contidas na ontologia possam realizar suas estimativas (5). Depois de extraídos, eles são instanciados na máquina de regras (6), usando a ontologia de sintonia fina (7). A partir daí, através das heurísticas de sintonia fina, definidas na ontologia (5), a máquina de regras (6) infere as ações de sintonia fina (8), dada aquela carga de trabalho (3). Caso seja o modo automático (9), serão executadas todas as ações inferidas no banco de dados (1). Já se for um processo semiautomático, as ações de sintonia fina deverão ser exibidas pela interface ao usuário (10), que escolherá (11) quais ações executará no banco de dados (1).

4.3 Projeto de arquitetura baseada em componentes

Todo o *design* de um *framework* é criado a partir de como as funcionalidades vão ser transformadas em código fonte e como elas vão se comunicar entre si. Durante o projeto de arquitetura foi realizado o planejamento dos passos necessários para a inferência de ações de sintonia fina (Figura 1). A partir daí, criou-se o projeto de arquitetura do *framework* Outer-Tuning (Figura 2), que foi seguido de sua implementação.

A arquitetura escolhida para o Outer-Tuning é baseada em componentes. Componente de *software* é uma parte não-trivial, quase independente, e substituível, de um sistema que cumpre uma função clara no contexto da arquitetura [5]. Devido ao caráter experimental da ferramenta e as múltiplas tecnologias envolvidas (ex.: SGBDs, máquinas de regras, ontologia, bibliotecas, mais de uma linguagem de programação), decidiu-se que os componentes poderiam facilitar a comunicação entre as partes e deixar cada uma das fases de execução (Figura 1) independentes, e caso necessário, poderiam ser substituídos/mantidos de forma compartimentalizada sem a propagação de bugs. Além dessas vantagens, componentes são umas das mais viáveis e usadas formas de reuso de *software*, principalmente entre desenvolvedores de *software* livre [15].

Como resultado da escolha de uma arquitetura baseada em componentes, decidiu-se que o Outer-Tuning seria desenvolvido como um *framework* de aplicação, que por definição é uma aplicação semi-completa, construída com uma coleção

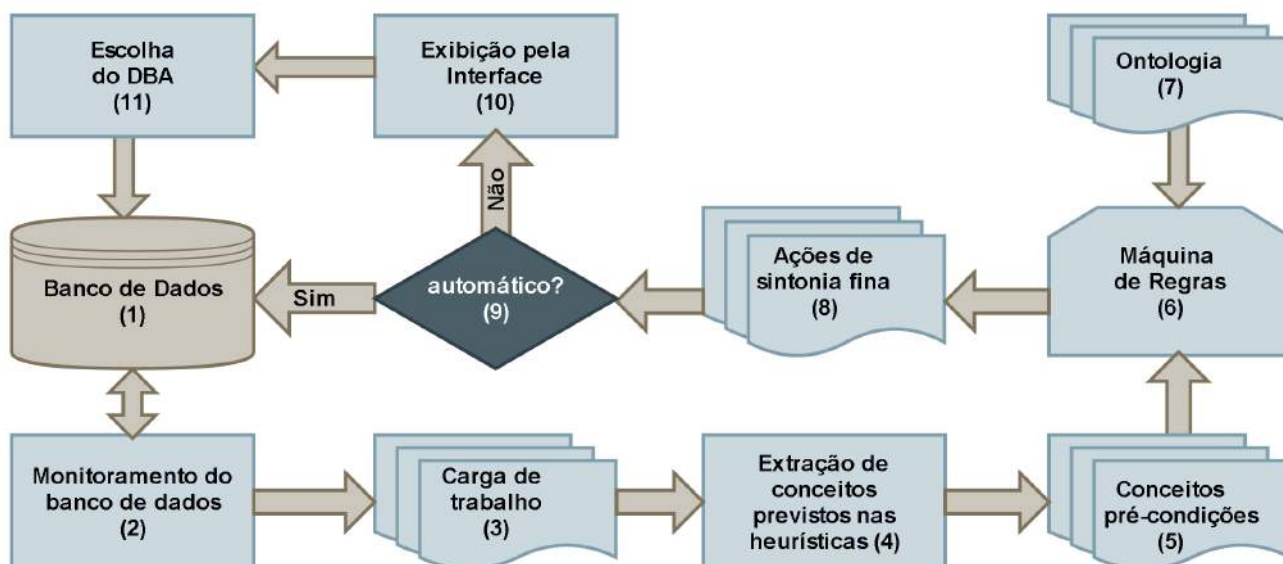


Figura 1: Fluxo de execução projetado para o *framework* Outer-Tuning

organizada de componentes de *software* reusáveis [16]. A escolha desse tipo de *framework*, assim como o uso de componentes, foi feita para dar qualidade ao *software*, possibilitar uma futura evolução para *software* orientado a serviços [3], mas principalmente permitir um baixo acoplamento entre as partes do *software*.

De acordo com o fluxo de execução proposto, o Outer-Tuning teve seus componentes definidos e especializados para cada etapa do ciclo. Na Figura 2, são enumerados e descritos de forma detalhada os principais elementos da arquitetura proposta:

(1) **Base:** Apesar do desenvolvimento baseado em componentes ser uma das formas de reuso de *software* mais eficientes da engenharia de *software*, foi utilizada em paralelo a esse tipo de reuso uma biblioteca de funções para que os componentes pudessem compartilhar funções ordinárias e redundantes, além de dar acesso a um sistema de log compartilhado.

(2) **Ontologia:** A ontologia de aplicação contém a ontologia de domínio (define conceitos) e a ontologia de tarefas (define as heurísticas através de regras “se-então”). É a principal parte extensível do *framework*. Durante o projeto, ela foi propositalmente isolada da implementação para que o DBA pudesse inserir/alterar/excluir heurísticas sem ter que alterar componentes do *framework*. Na prática isso dá ao DBA uma independência da implementação, ele não precisa saber como o *framework* foi implementado para, por exemplo, corrigir um modelo de custo de alguma heurística. As regras são definidas em uma linguagem declarativa (SWRL), logo o DBA precisa se preocupar apenas em “o que” a heurística deve fazer, e não “como” ela deve fazer.

(3) **Máquina de regras:** Uma das decisões importantes tomadas, foi decidir “em qual momento” e “como” a máquina de regras seria utilizada. Uma máquina de regras (ou motor de inferência) é o componente pelo qual as regras “se-então” das heurísticas (definidas na ontologia de tarefas) são selecionadas e executadas. Foi usado um componente do tipo “caixa-preta”, onde não é possível o acesso ao código-fonte.

A única maneira de adaptar o componente é substituí-lo por outro que possua interfaces compatíveis com o atual. Para a implementação foi utilizada a máquina de regras Jess². Sua escolha se deu pela compatibilidade entre as interfaces e os requisitos. O requisito funcional (iv) (Seção 4.1) determina que o DBA possua uma forma de habilitar e desabilitar as heurísticas antes de iniciar um processo de sintonia fina. A biblioteca Jess possui uma interface que permite habilitar/desabilitar um conjunto de regras SWRL (usadas nas heurísticas), e isso foi um aspecto decisivo para que o DBA, através da interface do Outer-Tuning, pudesse escolher quais heurísticas deveriam ser executadas ou não.

(4) **Banco de dados:** Este também foi considerado um componente caixa preta. Sua comunicação com o *framework* (5) e (15) se dá através dos drivers de conexão usados pelos componentes *CapturadorCargaDeTrabalho* (6) e *ExecutorDeFuncoes* (7). Vale lembrar que, o termo banco de dados aqui tem o significado de SGBD. Logo, a comunicação pode ser realizada com qualquer banco de dados gerenciado por um determinado SGBD.

(5) **Carga de trabalho:** A carga de trabalho capturada consiste em comandos SQL do tipo DML (*Data Manipulation Language* - Linguagem de manipulação de dados), os seus respectivos planos de execução e sua frequência. O termo carga de trabalho não abrange os dados manipulados por estes comandos DML.

(6) **CapturadorCargaDeTrabalho:** É o componente responsável por adquirir a carga de trabalho com um intervalo de tempo pré-determinado pelo DBA para realizar a sintonia fina. Para a conexão entre o banco de dados e o *framework* foi projetada uma classe de conexão que utiliza o driver JDBC³. Logo, o Outer-Tuning pode suportar os principais SGBDs do mercado que implementam essa tecnologia.

(7) **ExecutorDeFuncoes:** Toda heurística possui um conjunto de conceitos que são pré-condições e que devem ser

²<http://www.jessrules.com/> acesso em 18/02/2015

³<http://www.oracle.com/> acesso em 18/02/2015

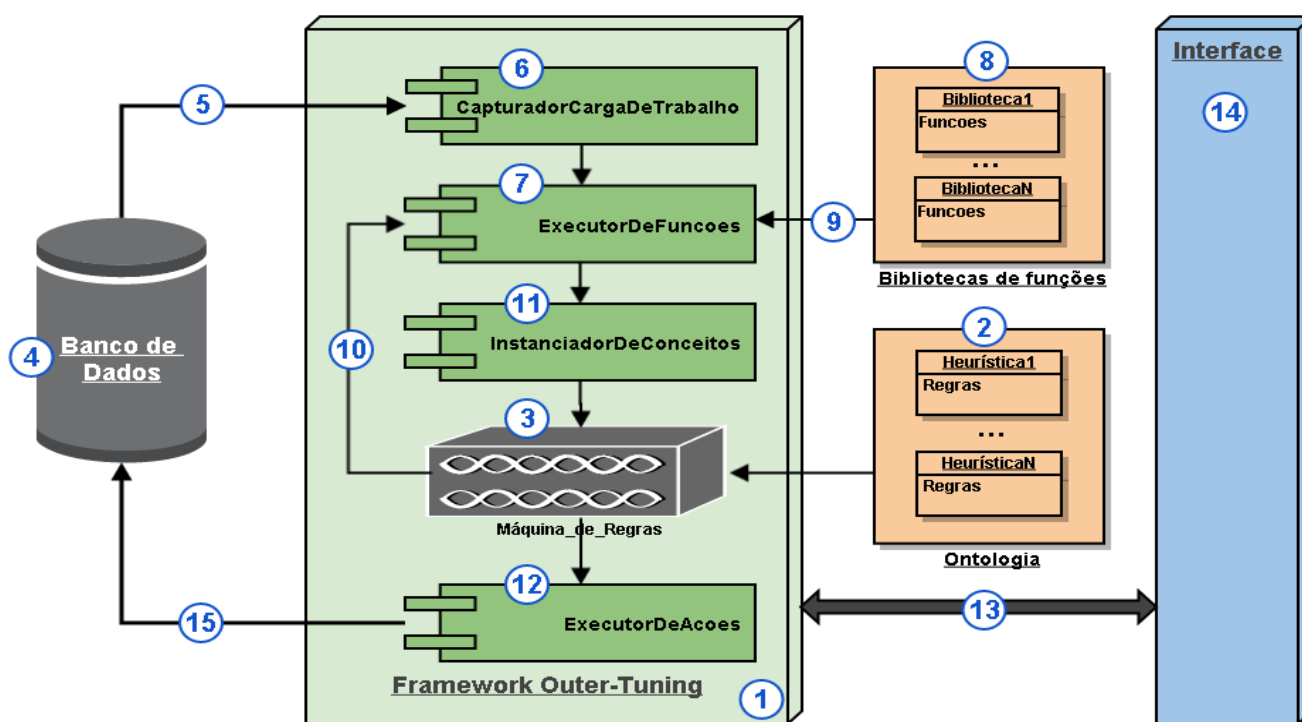


Figura 2: Projeto de Arquitetura do *framework* Outer-Tuning

instanciados na máquina de regras para que haja inferência de ações de sintonia fina. O componente *ExecutorDeFuncoes* é o responsável por extrair informações da carga de trabalho e gerar os indivíduos de conceitos que são pré-condições. Ele possui três fontes de entrada de dados para este componente:

(i) Carga de trabalho capturada pelo componente *CapturadorCargaDeTrabalho* (6); (ii) Definição das funções, parâmetros de entrada e saída (10) que são definidas pelas heurísticas escolhidas pelo DBA durante a inicialização da ferramenta, para instanciar os conceitos que são seus pré-requisitos. (iii) Implementação das funções (9) obtidas pelas bibliotecas de funções implementadas e compiladas. Essas bibliotecas são fisicamente independentes do *framework* e lidas em tempo de execução pelo componente *ExecutorDeFuncoes* (7). Isso faz com que elas possam ser adicionadas ou substituídas sem nenhuma intervenção no código ou (re)compilação da ferramenta.

Para realizar o seu trabalho, o *ExecutorDeFuncoes* recebe do componente *CapturadorCargaDeTrabalho* (6) a carga de trabalho capturada. Também recebe da máquina de regras (10) a lista de conceitos que devem ser extraídos, junto com as assinaturas das funções que realizaram a extração. De posse das assinaturas das funções, o componente *ExecutorDeFuncoes* lê (9) a biblioteca que contém o código fonte compilado das funções, e as executa. O parâmetro de entrada das funções é sempre originado da carga de trabalho e os parâmetros de saída são sempre indivíduos de conceitos pré-condições das heurísticas. De posse dos indivíduos, resultado da execução das funções, o componente *ExecutorDeFuncoes* envia-os para o componente *InstanciadorDeConceitos*.

(8) **Bibliotecas de funções:** Sua função é aglutinar bibliotecas de código fonte compiladas responsáveis por extrair

conceitos da carga de trabalho. As bibliotecas de funções, assim como a ontologia, são partes extensíveis do *framework* e foram projetadas para trabalhar isoladas da estrutura do *framework*. Elas são lidas e executadas em tempo de execução, podendo ser incluídas, alteradas, ou apagadas sem nenhuma intervenção no código fonte do *framework*.

(9) **Comunicação entre bibliotecas e *ExecutorDeFuncoes*:** É realizada através de interface definida no componente *ExecutorDeFuncoes* que busca no repositório as funções desejadas.

(10) **Conceitos pré-condições das heurísticas:** O componente *ExecutorDeFuncoes* recebe da máquina de regras os conceitos que são pré-condições e a assinatura das funções contidas na biblioteca de funções que extraem conceitos.

(11) *InstanciadorDeConceitos*: É responsável por instanciar os indivíduos gerados pela execução das funções pelo *ExecutorDeFuncoes* na máquina de regras. Estes indivíduos são de conceitos que são pré-condições para execução das heurísticas. Foi determinado que esta funcionalidade se tornasse um componente pela complexidade da tarefa. Este componente tem a responsabilidade de instanciar qualquer conceito, atributo ou relacionamento de conceitos definidos na ontologia, independente da versão da ontologia ou heurística escolhida. Ele também é responsável por não instanciar dois indivíduos iguais, visto que as bibliotecas de funções podem extrair conceitos iguais em momentos diferentes. Por exemplo: duas consultas podem compartilhar uma mesma tabela “vendas”. Logo, este componente controla indivíduos duplicados durante a instanciação.

(12) **ExecutorDeAcoes**: Monitora a máquina de regras, e captura as ações de sintonia fina inferidas pelas heurís-

ticas. Após a captura, o componente transmite as ações de sintonia fina para a interface do *framework* para que o DBA possa escolher quais ações deseja efetivamente executar. Após a escolha, o componente *ExecutorDeAcoes* recebe uma lista de ações da interface e as executa no banco de dados. Compartilha o sistema de conexão com o componente *CapturadorCargaDeTrabalho* através da *Base* (1).

(13) Comunicação *framework* – interface: Por se tratar de uma ferramenta de sintonia fina, considerou-se importante que a interface fosse desacoplada do *framework*. A independência é justificada pela possibilidade de dar ao DBA a opção de rodar a ferramenta em sistemas operacionais sem interface gráfica (ex.: servidores Linux) que proveem potência computacional, e ao mesmo tempo poder usufruir de uma interface amigável e que possa evoluir sem nenhuma restrição tecnológica em relação ao Outer-Tuning. Para que isso fosse viável, decidiu-se implantar um protocolo de comunicação entre a Interface(14) e o *framework*. O padrão de projeto escolhido foi o quadro-negro (*blackboard*)[11], pela sua facilidade de implementação, e pela possibilidade da execução sem nenhuma interface. Com a utilização do padrão quadro-negro, a interface durante os testes e implementação pôde ser substituída por um arquivo de configuração, que facilita a depuração do *framework* de forma mais rápida e prática por parte do desenvolvedor.

(14) Interface: É responsável pela interação com o DBA. Foi projetada para ser uma interface web e possuir 5 seções: (a) Seleção de heurísticas: possibilita o usuário selecionar quais heurísticas deseja executar; (b) carga de trabalho capturada: é possível visualizar estatísticas e a carga de trabalho executada no banco de dados em tempo real; (c) ações de *tuning*: permite selecionar e executar as ações inferidas pela ontologia no banco de dados; (d) ações hipotéticas: exibe as ações de sintonia fina hipotéticas, que ainda não foram consideradas positivas para a carga de trabalho mas são opções possíveis de sintonia fina; e (e) log: tela de visualização do log de execução do *framework*. Permite ao DBA observar o comportamento da ferramenta e detalhes internos de sua execução.

4.4 O uso de linguagens lógicas para realizar sintonia fina

Linguagens lógicas têm sido usadas com frequência como base para a construção de modelos e ferramentas para representar conhecimento. A Lógica Clássica de Primeira Ordem (LPO) poderia ser uma escolha natural mas é, no entanto, indecidível, ou não possui mecanismo para verificação computacional de consistência de uma base de conhecimento.

Description Logic (DL), por outro lado, representa um fragmento decidível de LPO. O poder de expressão de DL é bastante abrangente, permitindo desde a representação de hierarquias simples até modelos associados a diagramas de classes e/ou comportamento em UML, passando por modelos de Entidades e Relacionamentos. DL é, de certa forma, uma herança de iniciativas como DATALOG, que permitem a especificação e validação de propriedades impossíveis de serem verificadas em LPO. Em DATALOG é possível se operar queries recursivas sem a necessidade de definir transações. Pelas razões teóricas descritas acima, em conjunção com outras razões históricas, DL tornou-se a lógica subjacente na formalização da maior parte das Ontologias que necessitam de algum teste formal de validação.

Existem diversas lógicas de descrição com diferentes capa-

cidades de especificação. A OWL-DL (Ontology Web Language - DL) é a lógica de descrição que foi escolhida para implementar o aspecto dedutivo da Web Semântica. Diversas ontologias na Web são formalizadas em OWL-DL. Trata-se de uma linguagem que tem um poder de expressão bastante significativo, sendo possível a garantia de representação de qualquer modelo de classes em UML e/ou ER, com verificação automática de consistência [4].

Neste nosso trabalho, o uso de regras para representar as heurísticas de sintonia fina em bancos de dados relacionais se mostra mais natural que puramente uma descrição via ontologia das mesmas. Por esta razão, a adição de regras à OWL-DL mostrou-se bastante atrativa. Dentre as diversas possibilidades, optamos por adicionar regras DATALOG, a saber, rulesML à OWL-DL. Na literatura, SWRL (Semantic Web Rules Language) é precisamente o que mais se adequa à nossa necessidade.

O trabalho que descrevemos neste artigo resolve o problema da integração de regras à OWL-DL implementando o mapeamento de regras restritas SWRL em DL [10]. O único cuidado que tomamos que garante a adequação do mapeamento, é que em todas as regras de sintonia fina não há variáveis compartilhadas entre antecedentes de uma regra e consequentes da mesma regra, e no mínimo um nominal é compartilhado entre antecedente e consequente. Desta forma, a nossa implementação de SWRL para sintonia fina de Banco de Dados via ontologias em OWL-DL, apesar do uso de regras, é decidível.

5. AVALIAÇÃO

Faz parte da definição de um *framework*, a capacidade de ser estendido. Como método de avaliação, o Outer-Tuning foi estendido para a inserção de duas novas heurísticas de sintonia fina, ambas responsáveis pela seleção e criação de visões materializadas: a heurística de benefícios [14] e a heurística de expectativas [17]

Devido sua arquitetura, todos os componentes (com exceção dos considerados “caixa-preta”) do Outer-Tuning são passíveis de extensão. No caso de inserção/alteração/exclusão de heurísticas de sintonia fina no Outer-Tuning, são necessárias apenas alterações no arquivo da ontologia, e se necessário, inclusão de novas funções para extração de conceitos nas bibliotecas de funções (veja (2) e (8) Figura 2). Essa característica tornou o processo de inclusão de duas novas heurísticas de VMs um processo rápido. Foram necessárias apenas o estudo das heurísticas e a formalização de suas regras de inferência na linguagem SWRL. Não foi necessário alteração de código fonte ou conhecimento da implementação. Todas as alterações se mantiveram em alto nível, com manipulação apenas da ontologia.

As heurísticas foram executadas e tiveram seus resultados avaliados e comparados através do Outer-Tuning. Para a geração da carga de trabalho durante os testes foi utilizado o *benchmark* TPC-H que possui uma carga de trabalho analítica (OLAP), propícia para a avaliação de ferramentas de seleção de visões materializadas. A carga de trabalho foi utilizada com o SGBD PostgreSQL na versão 9.1 e sistema operacional Ubuntu 12.04 32bits. O hardware usado foi um Quad-Core 1.6Ghz, 4GB de RAM e 100GB de disco rígido.

Foram considerados dois tipos de VMs durante a análise dos resultados, as benéficas e as maléficas. Uma VM foi considerada benéfica quando as consultas reescritas e forçadas

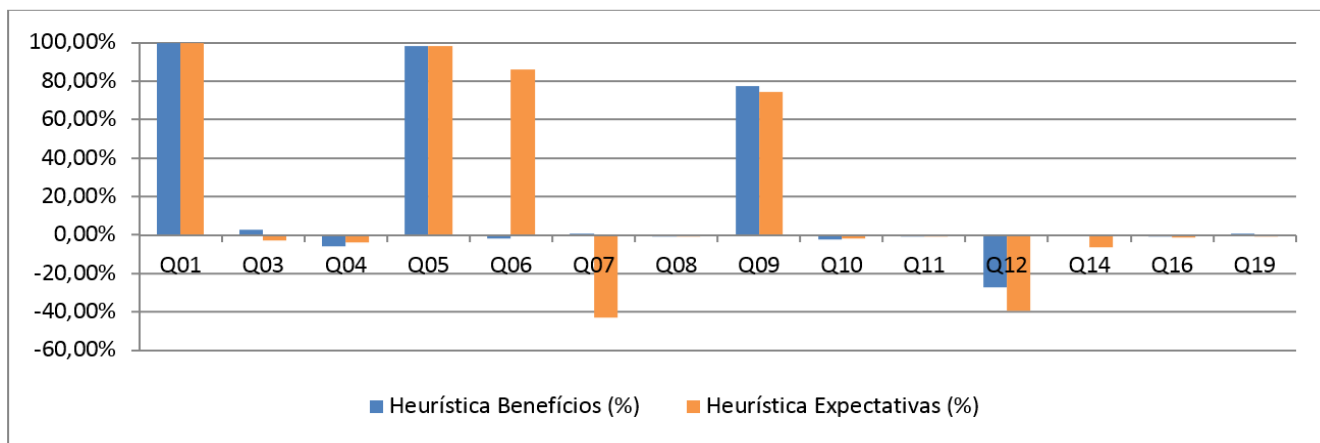


Figura 3: Ganho (em %) no tempo de execução das consultas na presença de visões materializadas sugeridas pelas heurísticas de benefícios e de expectativas

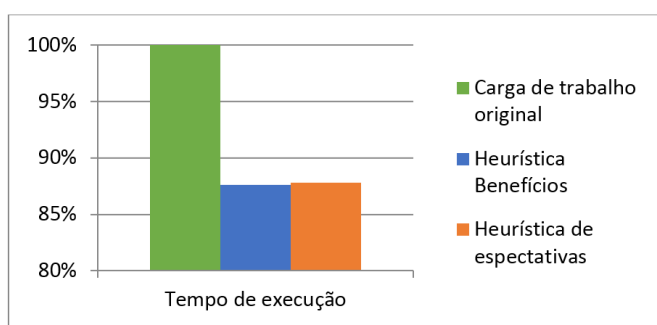


Figura 4: Custo total de cada heurística em relação à carga de trabalho original

a utilizarem a VM trouxeram um ganho para a execução da carga de trabalho, e malélicas quando trouxeram perdas para a execução da carga de trabalho. Uma VM malélica ocorre quando a reescrita de consulta remove parâmetros de restrição e torna a VM maior que a quantidade de páginas lidas na consulta original. Mais detalhes sobre VMs malélicas em [17].

As duas heurísticas inferiram visões materializadas para a carga de trabalho utilizada. A heurística de benefícios apontou três visões benéficas para a carga de trabalho (Q01, Q05, Q09) e duas malélicas (Q4, Q12). Já a heurística de expectativa apontou quatro visões benéficas (Q01, Q05, Q06, Q09) e três malélicas (Q3, Q04, Q07, Q12, Q14) (Figura 3).

Ambas as heurísticas (Figura 4) trouxeram um ganho positivo equivalente para a carga de trabalho (12,4% e 12,2%). Porém o Outer-Tuning evidenciou que a heurística de benefícios teve o custo de criação e armazenamento de VMs menor que a heurística de expectativas. Enquanto a heurística de benefícios teve 5 VMs criadas, a de expectativas teve 9 VMs criadas.

Pôde-se comprovar, através do Outer-Tuning, que uma mesma heurística pode trazer benefícios ou prejuízos de acordo com a carga de trabalho utilizada. Entendeu-se que a qualidade das ações de sintonia fina está relacionada diretamente com a qualidade da heurística e com a compatibili-

dade entre a heurística e a carga de trabalho. Independentemente do resultado apresentado por ambas heurísticas, o teste mostra que o *framework* é capaz de trabalhar simultaneamente com mais de uma heurística, é capaz de comparar as soluções apresentadas e a inclusão de novas heurísticas é feita em alto nível através da ontologia.

Com o uso do Outer-Tuning, o DBA possui informações suficientes para avaliar quais heurísticas são interessantes para a sua carga de trabalho. Ele também consegue utilizar sua experiência para adequar as heurísticas através da alteração das mesmas na ontologia, e da inclusão de novas heurísticas. Nada impede, por exemplo, que o DBA realize uma junção entre as heurísticas de benefícios e expectativas criando a sua própria heurística capaz de selecionar as VMs positivas que ambas inferiram, mas evitando criar as VMs negativas (Q7, Q3, Q14) como fez a heurística de expectativas.

Foi mostrado pelos testes que o Outer-Tuning é capaz de: i) inferir ações de sintonia fina úteis; ii) comparar heurísticas de sintonia fina inseridas na ontologia; e iii) apoiar o DBA na tarefa de sintonia fina ao fornecer informações relevantes para que ele possa adequar as heurísticas à carga de trabalho ou inserir novas heurísticas na ontologia de sintonia fina.

Para que o leitor possa conhecer detalhes do funcionamento, visualizar as informações que são disponibilizadas ao DBA e a dinâmica do processo de sintonia fina, está disponível no endereço <http://outertuning.biobd.inf.puc-rio.br> uma versão demonstrativa.

6. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foi apresentado o fluxo de controle e a arquitetura usada para desenvolver uma ferramenta de apoio à sintonia fina de banco de dados. Essa ferramenta foi projetada de forma a possibilitar a sua extensão em alto nível, sem exigir que o DBA precise demandar tempo para entender o código-fonte. Além disso, foi usado um *benchmark* conhecido pela comunidade de banco de dados e foram aplicadas duas heurísticas diferentes para validar que a ferramenta consegue alternar entre heurísticas de forma

amigável, ou seja, em alto nível, usando apenas a ontologia e componentes independentes da forma de implementação. Dessa forma, foi possível comparar os resultados de ambas heurísticas, mesmo que estes se mostrassem com ganhos parecidos. O objetivo do presente trabalho não é demonstrar que uma heurística é ou não melhor do que outra, mas sim possibilitar a flexibilidade ao DBA na escolha, comparação e extensibilidade de uso de heurísticas diferentes. E durante essa comparação, obter justificativas semânticas para as decisões tomadas pelas heurísticas descritas na ontologia usada pela ferramenta.

Como trabalho futuro se pretende elaborar um template para padronização e melhoria da apresentação das justificativas semânticas e apresentar a ontologia de forma gráfica para que o DBA possa manipular os conceitos que ele precisa indicar como necessários ao inserir uma nova heurística. Além disso, encontrar uma maneira gráfica de demonstrar toda a evolução do raciocínio da ferramenta para a tomada de determinada decisão.

Adicionalmente, nossa ferramenta se limitou a cargas de dados estáticas do tipo OLAP (*On-line Analytical Processing*) com o objetivo de manipular e analisar um grande volume de dados sob múltiplas perspectivas, e não considerou o custo de manutenção das visões materializadas durante o processo de seleção das ações de sintonia fina. Contudo, somente será necessária a adequação das heurísticas e da ontologia para que o Outer-Tuning possa considerar os custos de atualização dos dados durante o processo de *tuning*. Logo, como trabalho futuro, esperamos remover essa limitação e selecionar ações de sintonia fina para outros tipos de carga de trabalho.

7. REFERÊNCIAS

- [1] A. C. B. d. Almeida. *Framework para apoiar a sintonia fina de banco de dados*. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro - PUC-RIO, 2013.
- [2] K. Aouiche, P.-E. Jouve, and J. Darmont. Clustering-based materialized view selection in data warehouses. In *10th E.E. Conf. on Advances in Databases and Information Systems*, ADBIS'06, pages 81–95, Berlin, Heidelberg, 2006. Springer-Verlag.
- [3] L. G. Azevedo, F. A. Baião, and F. Santoro. Identificação de Serviços a partir da Modelagem de Processos de Negócio. *V Simpósio Brasileiro de Sistemas de Informação*, pages 133–144, 2009.
- [4] D. Berardi, D. Calvanese, and G. De Giacomo. Reasoning on UML class diagrams. *Artificial Intelligence*, 168:70–118, 2005.
- [5] A. W. Brown and K. C. Wallnau. Engineering of Component Based Systems. In *Component-Based Software Engineering*, pages 7–15, 1996.
- [6] N. Bruno. *Automated Physical Database Design and Tuning*. CRC Press, New York, New York, USA, 2012.
- [7] A. W. Carvalho. Criação Automática de Visões Materializadas em SGBDs Relacionais. Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro - PUC-RIO, 2011.
- [8] G. Chan, Q. Li, and L. Feng. Optimized design of materialized views in a real-life data warehousing environment. *International Journal of Information Technology*, 7(1):30–54, 2001.
- [9] R. Chirkova and J. Yang. Materialized views. *Found. and Trends in Databases*, pages 295–405, 2012.
- [10] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosz, and M. Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, 2015.
- [11] R. Khosla, N. Ichalkaranje, and L. C. Jain. *Design of Intelligent Multi-Agent Systems: Human-Centredness, Architectures, Learning and Adaptation*. Studies in Fuzziness and Soft Computing. Springer, 2004.
- [12] M. Lawrence. Multiobjective genetic algorithms for materialized view selection in OLAP data warehouses. In *GECCO '06*, page 699. ACM Press, 2006.
- [13] X. Li, X. Qian, J. Jiang, and Z. Wang. Shuffled Frog Leaping Algorithm for Materialized Views Selection. *Intl. Workshop Education Technology and Computer Science*, pages 7–10, 2010.
- [14] E. M. T. Morelli and S. Lifschitz. Recriação Automática de Índices em um SGBD Relacional. Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro - PUC-RIO, 2006.
- [15] L. D. C. Nascimento and F. Santoro. Análise de interações nas Comunidades Virtuais de Software Livre. *V Simpósio Brasileiro de Sistemas de Informação*, pages 12–23, 2009.
- [16] J. L. D. Oliveira, L. Fernando, B. Loja, S. Larissa, V. Vicente, and G. Neto. Um Componente para Gerência de Processos de Negócio em Sistemas de Informação. *VII Simpósio Brasileiro de Sistemas de Informação*, pages 250–261, 2011.
- [17] R. P. d. Oliveira. Sintonia fina baseada em ontologias: o caso das visões materializadas. Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro - PUC-RIO, 2015.
- [18] W. Pree. Hot-spot-driven framework development. In *Summer School on Reusable Architectures in Object-Oriented software Development*, pages 123–127. ACM, 1995.
- [19] R. Ramakrishnan and J. Gehrke. *Sistemas de gerenciamento de banco de dados*. McGraw Hill, 2008.
- [20] D. Shasha and P. Bonnet. *Database Tuning: Principles, Experiments, and Troubleshooting Techniques*. Elsevier Science, 2002.
- [21] X. Song and L. Gao. An ant colony based algorithm for optimal selection of materialized view. In *Intelligent Computing and Integrated Systems (ICISS)*, pages 534–536, Oct 2010.
- [22] X. Sun and Z. Wang. An Efficient Materialized Views Selection Algorithm Based on PSO. In *Intl. Intelligent Systems and Applications*, pages 1–4. Ieee, May 2009.
- [23] S. H. Talebian and S. A. Kareem. A lexicographic ordering genetic algorithm for solving multi-objective view selection problem. In *Conf. on Comp. Research and Development*, ICCRD '10, pages 110–115, 2010.
- [24] T. Vijay Kumar and S. Kumar. Materialized view selection using genetic algorithm. In *Contemporary Computing*, volume 306, pages 225–237. Springer Berlin Heidelberg, 2012.
- [25] Z. Yuhang, L. Qi, and Y. Wei. Materialized view selection algorithm cssa vsp. In *Computational Intelligence and Natural Computing Proceedings (CINCP)*, volume 1, pages 68–71, Sept 2010.

Aplicação de Descoberta de Conhecimento em Bases de Dados na Estimativa da Evapotranspiração: um Experimento no Estado do Rio de Janeiro

Alternate Title: Application of Knowledge Discovery in Databases in Evapotranspiration Estimation: an Experiment in the State of Rio de Janeiro

Fernando Xavier
Universidade Federal do Estado do
Rio de Janeiro (UNIRIO)
Avenida Pasteur, 458 - Urca
Rio de Janeiro – RJ – Brasil
fernando.xavier@uniriotec.br

Asterio Kiyoshi Tanaka
Universidade Federal do Estado do
Rio de Janeiro (UNIRIO)
Avenida Pasteur, 458 - Urca
Rio de Janeiro – RJ – Brasil
tanaka@uniriotec.br

Kate Cerqueira Revoredo
Universidade Federal do Estado do
Rio de Janeiro (UNIRIO)
Avenida Pasteur, 458 - Urca
Rio de Janeiro – RJ – Brasil
katerevored@uniriotec.br

RESUMO

Com o crescimento do volume de dados em diversas áreas, como a Hidrologia, aumenta a necessidade do uso de sistemas de informação para auxílio na manipulação desses dados. Este artigo é um relato de um experimento que usou técnicas de descoberta de conhecimento para a estimativa de um importante componente do ciclo hidrológico: a evapotranspiração. O experimento relatado neste artigo foi realizado com dados meteorológicos e mostrou que alguns algoritmos, como o M5P, apresentam bons resultados quando comparados com os dados históricos da evapotranspiração estimada.

Palavras-Chave

KDD, Mineração de Dados, Hidrologia, Evapotranspiração, Regressão Linear.

ABSTRACT

With the growing volume of data in various areas such as Hydrology, there is a need for using information systems to aid in handling such data. This article is a report of an experiment that used knowledge discovery techniques to estimate an important component of the hydrological cycle: evapotranspiration. The experiment reported in this article was done with weather data and showed that some algorithms, such as M5P, present good results when compared to historical data of the estimated evapotranspiration.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.

Copyright SBC 2015.

Categories and Subject Descriptors

H.2.8 [Information Systems]: Database Applications – data mining.

General Terms

Algorithms, Experimentation, Theory.

Keywords

KDD, Data Mining, Hydrology, Evapotranspiration, Linear Regression.

1. INTRODUÇÃO

A água é um recurso fundamental para o planeta e, por conta disso, existem diversas iniciativas destinadas a estudar os assuntos relacionados a este recurso. Uma dessas iniciativas é o programa Intergovernamental Panel on Climate Change (IPCC), mantido pela Organização das Nações Unidas (ONU) e pela Organização Meteorológica Mundial (WMO), para avaliar as informações científicas, técnicas e socioeconômicas produzidas no mundo, que são importantes para a compreensão das mudanças climáticas [15], que está relacionada a interesses em atividades como: monitoramento de desastres naturais, agricultura, abastecimento de água, geração de energia, entre outros.

Esses interesses são objetos de estudos de ciências como a Hidrologia, que é a ciência que trata da água no planeta, sua ocorrência, circulação e distribuição [3]. As pesquisas relacionadas à Hidrologia usam, por exemplo, dados de séries históricas de chuvas, vazões dos rios, umidade do ar, temperatura, dentre outros, que são importantes para a compreensão dos fenômenos naturais, como o ciclo hidrológico, que tem influência no clima de diversas formas [3].

Dentre os componentes do ciclo hidrológico, destaca-se a evapotranspiração, que é a soma da evaporação da água do solo com a transpiração da vegetação, que retorna para a atmosfera na forma de vapor [2]. Devido ao fato da medida direta da evapotranspiração ser difícil e onerosa, a abordagem corrente é

sua estimativa através de métodos matemáticos [2], como a equação de Penman-Monteith, que é atualmente usada como referência pela Organização das Nações Unidas para Alimentação e Agricultura (FAO) [5]. Uma limitação dessa abordagem é que ela depende da disponibilidade dos valores de todas as variáveis da equação, o que impede que ela seja usada na ausência de valor de uma delas.

Tendo em vista essa limitação, o presente trabalho busca responder à seguinte questão: é possível estimar a evapotranspiração independentemente da disponibilidade de todas as variáveis? Para responder a esta questão, propõe-se uma alternativa para que a estimativa seja feita por modelos gerados através do processo de descoberta de conhecimento em banco de dados (em inglês Knowledge Discovery in Databases - KDD) [6]. Usando *datasets* com séries históricas de dados meteorológicos, espera-se, através de KDD, gerar um modelo que possa estimar o valor da evapotranspiração mesmo que os valores de todas as variáveis do *datasets* não estejam disponíveis.

Para avaliar a solução proposta, os valores calculados no modelo gerado são comparados com os valores históricos, disponíveis nos *datasets* utilizados no experimento executado neste trabalho. Se houver uma boa aproximação entre os valores gerados pelos modelos e os valores das séries históricas, então se pode demonstrar a viabilidade desta abordagem para a estimativa da evapotranspiração.

O restante do artigo está organizado como se segue. Na Seção 2, é apresentada a fundamentação teórica dos conceitos relacionados a este trabalho. Já na Seção 3, é descrita a metodologia utilizada neste trabalho, seguida na Seção 4 pela descrição da execução do experimento. Na Seção 5, é feita uma análise dos resultados encontrados e, na Seção 6, é apresentado um estudo dos trabalhos relacionados indicando as contribuições deste trabalho. Por fim, são feitas as considerações finais a respeito deste trabalho, indicando as contribuições esperadas, limitações e possíveis desdobramentos.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Evapotranspiração

A evapotranspiração, medida em milímetros/dia, é referida como a combinação dos processos de evaporação da água do solo e da transpiração da vegetação, que retorna à atmosfera em forma de vapor [2]. Conforme a Figura 1, a evapotranspiração está relacionada a outros componentes do ciclo hidrológico, como as chuvas, capacidade de infiltração do solo, dentre outros.

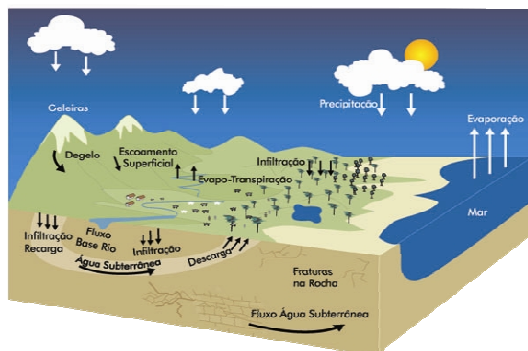


Figura 1. Ciclo Hidrológico [14]

Os valores da evapotranspiração são de fundamental importância em áreas como a agricultura, principalmente nas atividades de irrigação do solo. Nessas atividades, a captação e elevação da água são de grande importância, pois, além de consumirem energia e investimentos em equipamentos, podem ter impactos ambientais, como o uso inadequado dos recursos hídricos [18]. Nesse sentido, a disponibilidade do valor da evapotranspiração permite planejar melhor a quantidade de água a ser irrigada e, consequentemente, minimizar o uso dos recursos e os impactos ambientais.

A medida direta da evapotranspiração é difícil e onerosa, ao exigir instalações e equipamentos especiais, justificando seu uso apenas em condições experimentais [2]. Por isso, a abordagem corrente na área da Hidrologia é sua estimativa através de modelos matemáticos, a partir de variáveis meteorológicas, como a equação de Penman-Monteith [5], que é o método recomendado pela FAO para estimativa da evapotranspiração.

A equação de referência da FAO, como outros modelos matemáticos existentes, têm como dependência a disponibilidade dos valores das variáveis e, por isso, a estimativa da evapotranspiração é impossibilitada na ausência do valor de qualquer uma delas. Nesse sentido, o cálculo fica restrito apenas a áreas que dispõem de instrumentos de medição para todas as variáveis [16]. Em pequenas plantações, por exemplo, que dispõem de poucos recursos e têm a necessidade de planejar as atividades de irrigação do solo, a abordagem dos modelos matemáticos para estimativa da evapotranspiração torna-se inviável.

Outra limitação do método de Penman-Monteith é a estimativa da evapotranspiração apenas para áreas pequenas, sem aplicação para áreas maiores [15]. Visando superar essa limitação, existem métodos de estimativa da evapotranspiração que usam dados de sensoriamento remoto, como o método SEBAL [7].

2.2 Descoberta de Conhecimento em Banco de Dados

A Descoberta de Conhecimento em Bases de Dados é o processo de análise e interpretação dos dados para identificação de padrões compreensíveis, novos e potencialmente úteis, iterativo e composto pelas seguintes etapas (adaptadas de [6]), ilustradas na Figura 2:

- Pré-Processamento: etapa em que os dados são selecionados das bases de dados e uma sequência de transformações pode ser feita, como a seleção dos atributos que farão parte do conjunto de dados (*dataset*), limpeza dos dados, discretização dos valores, dentre outras.
- Mineração de Dados: Muitas vezes confundida com o processo de KDD em si, a mineração de dados permite aprendizado de modelos que reflitam os dados.
- Pós-Processamento: Avaliação e interpretação dos padrões que podem ser considerados novos conhecimentos [6].

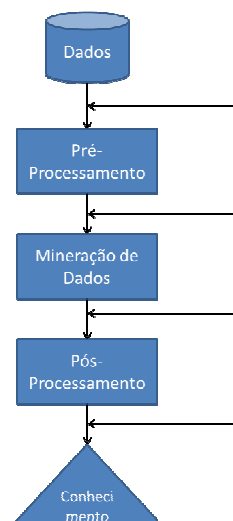


Figura 2. Etapas do Processo de KDD – Adaptado de [6]

Na mineração de dados, diversas técnicas podem ser usadas para aprendizado do modelo correspondente aos dados, seja para predição ou descrição:

- Classificação
- Regressão
- Análise de Séries Temporais
- Agrupamento
- Regras de Associação
- Caracterização

Essas técnicas são implementadas por algoritmos, como os baseados em redes neurais, árvores de decisão, regressão linear, dentre outros. O algoritmo M5P, por exemplo, é um algoritmo de classificação baseado em árvores de decisão [23], que aprende modelos para cálculo de valores que podem ser usados de acordo com um ponto de decisão, como o valor de algum atributo do *dataset*.

Esses algoritmos podem ser executados em ferramentas como o Weka [9] que, além da mineração de dados, automatiza as outras tarefas relativas ao processo de descoberta de conhecimento, além de possibilitar que a execução do processo seja repetida ou modificada usando diferentes configurações, como a estratégia de testes, uso de outros algoritmos, escolha da classe a ser analisada (no caso dos algoritmos de classificação), dentre outras.

O uso de KDD, ao automatizar a tarefa de descobrir um modelo que melhor se ajuste aos dados disponíveis, é uma abordagem que pode ser aplicada à estimativa da evapotranspiração, em contextos onde os dados que estão disponíveis podem ser bem diferentes de uma região para outra. Além disso, o uso de KDD é uma alternativa viável para aquelas regiões que não dispõem de todos os dados necessários da equação de referência da FAO para estimativa da evapotranspiração.

3. METODOLOGIA CIENTÍFICA

Segundo Recker [20], há a necessidade de se adquirir ao menos três tipos de conhecimento antes de se iniciar uma pesquisa científica:

- Conhecimento sobre o domínio ou tópico de interesse.

- Conhecimento sobre teorias relevantes sobre questões ou fenômenos.
- Conhecimento sobre os métodos de pesquisa que podem ser aplicados para se construir novos conhecimentos, artefatos ou novas questões.

Com base nisso, foram feitas revisões de trabalhos científicos sobre os assuntos relacionados ao tema de pesquisa deste trabalho. A partir dessas revisões, foram identificadas as principais questões relacionadas à estimativa da evapotranspiração, que foram usadas em novas revisões, de modo a identificar nos trabalhos relacionados possíveis problemas de pesquisa.

Com base nessa atividade, foi feita a formulação do problema de pesquisa, seguida da escolha, de acordo com os passos apresentados em [20], de qual estratégia de pesquisa a ser utilizada para se tentar solucionar o problema de pesquisa formulado. Neste trabalho, foi escolhido o método quantitativo, que pode ser definido como um conjunto de técnicas para responder questões de pesquisa com ênfase em dados quantitativos [20]. A razão para essa escolha se deve à natureza do problema de pesquisa, onde se deseja simplificar o cálculo de um componente do ciclo hidrológico.

Com a escolha do método, foi feito o planejamento da pesquisa, adaptando os procedimentos descritos por Recker [20] para esse tipo de método de pesquisa:

- Geração de modelos, teorias e hipóteses
- Desenvolvimento de instrumentos e métodos de medida
- Coleta de dados empíricos
- Modelagem Estatística
- Avaliação dos Resultados

A partir do problema de pesquisa, definiu-se a hipótese de que seria possível estimar a evapotranspiração, independentemente de quais dados estivessem disponíveis. Para teste da hipótese, definiu-se que seria executado o processo de KDD para avaliar os dados meteorológicos disponíveis na base de dados descrita na Seção 4.1 deste artigo.

Como ferramenta de apoio ao processo de KDD, utilizou-se o software Weka, que contém um conjunto de algoritmos de aprendizado de máquina para atividades de pré-processamento, mineração de dados e pós-processamento [9]. Com o auxílio de softwares como o Weka, pode-se executar facilmente para o mesmo *dataset* diversos algoritmos e executar análises como comparação dos resultados, gravação dos modelos gerados, importação de novos algoritmos, pré-processamento das instâncias, entre outras tarefas.

O experimento descrito neste trabalho se deu em três fases, relacionadas aos procedimentos descritos anteriormente e denominadas aqui de: exploração, execução e análise dos resultados. A fase de exploração compreendeu atividades como aprendizado do uso do Weka, análise das fontes de dados existentes, escolha dos atributos que iriam compor o *dataset* do experimento e uso de um *dataset* para análise dos resultados preliminares dos algoritmos de classificação. O objetivo desta fase era aprender sobre os atributos existentes nas fontes de dados e descobrir os algoritmos com melhores desempenhos para execução em outras estações de medição.

Já a fase de execução compreendeu a execução no Weka, do algoritmo escolhido na fase anterior, para mineração dos dados de *datasets* de estações de medição de diferentes regiões. O objetivo dessa fase foi gerar um modelo de estimativa da evapotranspiração para cada região analisada no estudo.

Na fase de análise, os resultados para cada região usada na mineração foram comparados com dados históricos e também com outros trabalhos realizados. O objetivo desta fase foi gerar conhecimento a partir da análise dos resultados deste experimento e discutir possíveis diferenças entre os resultados.

4. ESTIMATIVA DA EVAPOTRANSPIRAÇÃO ATRAVÉS DE KDD

4.1 Bases de Dados

Os dados foram extraídos do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET), que é um órgão responsável por prover informações meteorológicas, através do monitoramento, análise e previsão de tempo [13].

A escolha dessa base de dados se justifica porque contém séries históricas de dados meteorológicos para várias localidades do Brasil, coletados através de estações de medição. Além disso, desde 2006, essas séries históricas contém dados de evapotranspiração, que foram usados como dados de entrada para os algoritmos de aprendizado.

Para o experimento descrito neste trabalho, foram selecionadas todas as estações localizadas no Estado do Rio de Janeiro, descritas na Tabela 1. Conforme mostrado na Tabela 1, uma característica importante dessas estações é que não havia uniformidade na disponibilidade dos valores das variáveis, o que serviu para o objetivo do experimento de estimar a evapotranspiração independentemente de quais variáveis tinham valores nas séries históricas.

Após a seleção da base de dados e definição das estações de medição, foi feita a extração dos dados, selecionando-se na ferramenta disponibilizada pelo INMET todas as variáveis disponíveis, já que essa ferramenta apresenta uma lista de variáveis para as quais se deseja extrair a série histórica de valores.

Essa extração gerou *datasets* com as séries históricas de cada estação que, após a seleção das variáveis na extração, continham os seguintes atributos, com as definições trazidas por Branco [1] e INMET [13]:

- Velocidade Vento Média: Média das velocidades do vento do período.
- Velocidade Vento Máxima Média: Média das velocidades máximas do vento no período.
- Evaporação Piche: Medida da evaporação - em mililitro (ml) ou em milímetros de água evaporada - a partir de uma superfície porosa, mantida permanentemente umedecida por água.
- Evapotranspiração Potencial: Evapotranspiração estimada no mês.

- Insolação Total: Número de horas que a luz solar chegou até a superfície da Terra sem interferência de nuvens.
- Nebulosidade Média: A fração da abóbada celeste que é ocupada por nuvens.
- Precipitação Total: Quantidade de chuvas do período.
- Pressão Média: É a pressão exercida pelo peso da atmosfera em um determinado ponto da superfície da Terra.
- Temperatura Máxima Média: Média das temperaturas máximas do período.
- Temperatura Compensada Média: média entre as temperaturas máximas e mínimas, além de três medidas feitas ao longo do dia (9, 15 e 21 hs).
- Temperatura Mínima Média: Média das temperaturas mínimas do período.
- Umidade Relativa Média: Porcentagem de vapor d'água que há no ar.

Tabela 1. Estações de medição utilizadas no experimento

Estação	Dataset
Campos	Sem valores de pressão e insolação
Cordeiro	Sem valores de velocidade do vento
Itaperuna	Poucos valores de pressão
Paty do Alferes	Sem valores de pressão
Resende	Completo
Rio de Janeiro	Sem valores de insolação e poucos valores de pressão

4.2 Avaliação do Algoritmo

Na fase de exploração, foram selecionados para avaliação todos os algoritmos de classificação disponíveis no Weka, para execução em *datasets* que contém somente atributos numéricos.

As execuções dos algoritmos no Weka foram feitas utilizando o *dataset* da estação de Resende, por conter todas as variáveis, utilizando a estratégia de validação cruzada para particionamento dos conjuntos de treinamento e teste. Como avaliação complementar para seleção do algoritmo, utilizou-se a mesma abordagem usando-se o *dataset* da estação de Itaperuna. Em ambos os casos, o algoritmo de melhor coeficiente de correlação foi o M5P, conforme Tabela 2, que mostra os coeficientes por algoritmo para a estação de Resende:

Tabela 2. Coeficientes de correção para os principais algoritmos de classificação executados no Weka para o dataset da estação de Resende-RJ

Algoritmo	Coeficiente de Correlação
M5P	0.9893
MLPRegressor	0.9853
MultilayerPerceptron	0.9842
RBFRegressor	0.9840
Bagging	0.9812

O algoritmo M5P é uma adaptação feita por Wang e Witten [23] para o algoritmo M5, criado por Quinlan [19], que introduziu o conceito de árvores de modelos. Esse tipo de árvore é a combinação de uma árvore de decisão convencional com a possibilidade de ter, nas folhas da árvore, modelos lineares multivariados, como funções de regressão linear [19]. Na adaptação desse algoritmo para o M5P, foram incorporadas formas de tratar casos de instâncias com atributos enumerados ou sem valores, situações comuns em problemas do mundo real [23].

4.3 Aprendendo os Modelos Preditivos

Em seguida, após a definição do M5P como algoritmo a ser utilizado, foram feitas execuções do mesmo nos *datasets* das estações escolhidas para este experimento. A Tabela 3 mostra, para cada estação, as equações geradas pelo algoritmo:

Tabela 3. Modelos gerados pelo algoritmo M5P para cada estação

Estação	Decisão	Equação
Campos	-	$\text{EvapoBHPotencial} = 1.9879 * \text{NM} + 0.0377 * \text{PT} + 6.1418 * \text{TMaxM} + 2.2551 * \text{TCoM} + 5.7273 * \text{TMinM} - 0.7818 * \text{URM} - 198.7858$
Cordeiro	$\text{TCoM} \leq 20.252$	$\text{EvapoBHPotencial} = 0.1278 * \text{EP} + 0.0134 * \text{IT} + 0.0093 * \text{PT} - 0.6426 * \text{TMaxM} + 8.7045 * \text{TCoM} - 96.1925$
	$\text{TCoM} > 20.252$	$\text{EvapoBHPotencial} = 0.312 * \text{EP} + 0.0824 * \text{IT} + 1.5226 * \text{NM} + 0.0242 * \text{PT} - 5.0562 * \text{TMaxM} + 13.2391 * \text{TCoM} + 1.883 * \text{TMinM} - 133.0846$
Itaperuna	$\text{TCoM} \leq 23.585$	$\text{EvP} = 0.0381 * \text{EP} + 0.0829 * \text{IT} + 2.9742 * \text{NM} + 0.0362 * \text{PT} + 11.7057 * \text{TCoM} - 0.6848 * \text{TMinM} - 202.7576$
	$\text{TCoM} > 23.585$	$\text{EvP} = 0.1879 * \text{EP} + 0.214 * \text{IT} + 7.0959 * \text{NM} + 0.0062 * \text{PT} + 3.3705 * \text{TCoM} + 12.4856 * \text{TMinM} - 327.371$
Paty do Alferes	$\text{TCoM} \leq 19.635$	$\text{EvP} = 0.03 * \text{EP} + 2.0633 * \text{NM} + 0.0115 * \text{PT} + 0.5568 * \text{TMaxM} + 5.7038 * \text{TCoM} - 2.6609 * \text{TMinM} - 0.1132 * \text{URM} + 24.3532$
	$\text{TCoM} > 19.635$	$\text{EvP} = 0.1971 * \text{EP} + 0.4788 * \text{NM} + 0.0362 * \text{PT} - 1.4569 * \text{TMaxM} + 5.5637 * \text{TMinM} - 0.6512 * \text{URM} + 102.8731$
Resende	$\text{TCoM} \leq 20.276$	$\text{EvP} = 0.0388 * \text{EP} + 0.7079 * \text{NM} + 0.0068 * \text{PT} - 0.2215 * \text{PM} + 8.9358 * \text{TCoM} - 0.9358 * \text{TMinM} + 107.736$
	$\text{TCoM} > 20.276$	$\text{EvP} = 4.6224 * \text{VMM} + 0.0526 * \text{EP} + 2.6542 * \text{NM} + 0.0044 * \text{PT} - 0.8066 * \text{PM} + 11.8915 * \text{TCoM} - 0.609 * \text{TMinM} + 585.9188$
Rio de Janeiro	-	$\text{EvP} = 0.197 * \text{EP} + 4.7078 * \text{NM} + 5.9515 * \text{TMaxM} + 6.9072 * \text{TCoM} + 2.6752 * \text{TMinM} - 333.6851$

Onde as variáveis das equações geradas são descritas na legenda abaixo:

- EP: Evaporação Piche
- EvP: Evapotranspiração Potencial
- IT: Insolação Total
- NM: Nebulosidade Média
- PM: Pressão Média
- PT: Precipitação Total
- TCoM: Temperatura Compensada Média
- TMaxM: Temperatura Máxima Média
- TMinM: Temperatura Mínima Média
- URM: Umidade Relativa Média
- VMM: Velocidade do Vento Média

Nas figuras a seguir, são apresentados gráficos com os valores calculados pelo modelo gerado e os valores históricos dos *datasets* do INMET para cada uma das seis estações analisadas neste experimento. Para cada gráfico, executou-se a regressão linear para medir o grau de precisão entre os valores históricos e previstos, gerando-se o coeficiente R^2 e a função que relaciona os dois valores. O coeficiente de determinação, também chamado de R^2 , é uma medida de ajustamento de um modelo estatístico linear generalizado, como a regressão linear, em relação aos valores observados. O R^2 varia entre 0 e 1, indicando, em porcentagem, quanto o modelo consegue explicar os valores observados. Quanto maior o R^2 , mais explicativo é o modelo, melhor ele se ajusta à amostra [26].

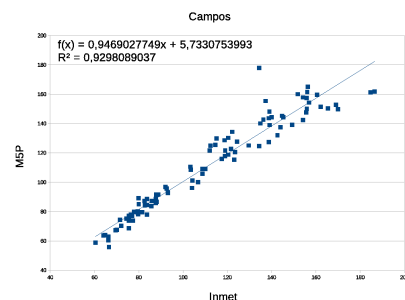


Figura 3. Evapotranspiração para a estação de Campos-RJ

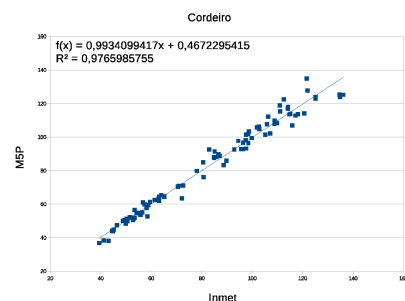


Figura 4. Evapotranspiração para a estação de Cordeiro-RJ

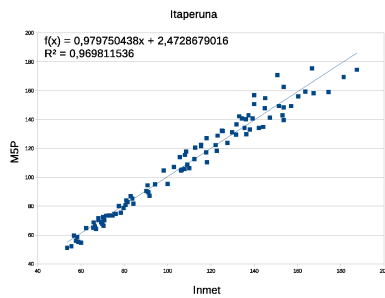


Figura 5. Evapotranspiração para a estação de Itaperuna-RJ

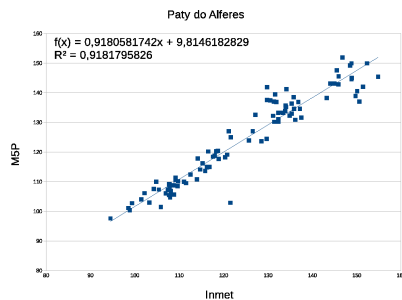


Figura 6. Evapotranspiração para a estação de Paty do Alferes-RJ

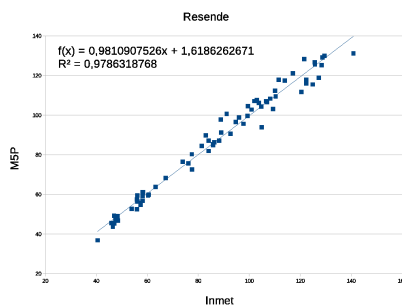


Figura 7. Evapotranspiração para a estação de Resende-RJ

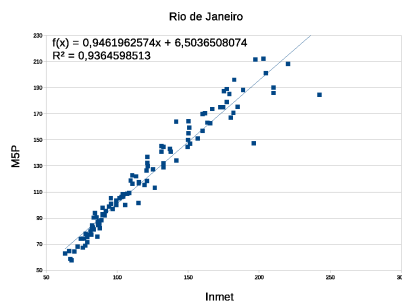


Figura 8. Evapotranspiração para a estação do Rio de Janeiro-RJ

5. ANÁLISE DOS RESULTADOS

A execução do algoritmo M5P gerou, como resultado, equações para cálculo da evapotranspiração potencial. Nas estações para as quais foram geradas duas equações, nota-se que a temperatura compensada média (TCoM) é o fator de decisão sobre qual equação utilizar. Isso demonstra a importância desse atributo para

cálculo da evapotranspiração potencial usando a abordagem proposta neste experimento.

Para duas estações (Campos e Rio de Janeiro), a execução do M5P gerou apenas uma equação. A característica dos *datasets* dessas estações em relação às demais é a ausência de valores em dois atributos: pressão e insolação. O *dataset* da estação do Rio de Janeiro continha apenas algumas instâncias com valores de pressão preenchidos.

Nota-se, na Tabela 3, que nenhuma das equações geradas usou todos os atributos. A explicação para esse fato é que o algoritmo M5P tem, como uma de suas características, a simplificação dos modelos lineares gerados, usando uma estratégia de busca gulosa para identificar e remover todos os atributos que contribuem pouco para o modelo, podendo até substituí-los por uma constante [19].

A única estação que continha todos os atributos preenchidos foi a de Resende, cujo valor de coeficiente de determinação (R^2) foi o mais alto, conforme a Figura 7. No entanto, os resultados para estação de Cordeiro apresentaram um desempenho semelhante, conforme a Figura 5. A característica do *dataset* dessa estação é a ausência de valores da velocidade do vento, conforme descrito na Tabela 1.

A semelhança entre os resultados para essas duas estações e o fato de que os valores de velocidade do vento só foram usados em uma única equação gerada entre todas as estações, é uma evidência de que essa variável pode não ter relevância no cálculo da evapotranspiração potencial, na abordagem utilizada neste experimento. Na equação de Penman-Monteith, a velocidade do vento é uma das variáveis requeridas, o que invalidaria o uso deste método para a estação de Cordeiro.

6. TRABALHOS RELACIONADOS

Em estudo relacionado à agricultura, Haghverdi et al [8] buscaram analisar comparativamente a mineração de dados com outras funções matemáticas conhecidas, para avaliar a produção de grãos baseados em dados de irrigação e evapotranspiração. Os autores identificaram que, nesse campo, os dados de evapotranspiração e salinidade do solo, que são difíceis de medir, podem ser substituídos por outras variáveis como quantidade da água irrigada e salinidade da água irrigada. Mostraram que a capacidade dos métodos de mineração utilizados (árvores de decisão e redes neurais) para introduzir novos preditores de entrada é uma vantagem, em relação aos tradicionalmente utilizados nesse campo.

Com o foco específico na estimativa da evapotranspiração, em [11] e [12] são descritas, entre outras coisas, a importância do uso de dados de evapotranspiração para o planejamento da irrigação de solos. Usaram processos gaussianos para entender a relação entre cada atributo e o valor da evapotranspiração. Entre seus resultados, mostraram que as redes neurais e processos gaussianos trouxeram maior acurácia do que os métodos de regressão linear.

Os trabalhos de Zhou et al [24] e Tang & Li [22] usaram o método Surface Energy Balance Algorithm for Land (SEBAL), que é baseado em dados de imagens de satélite, para comparação com outros métodos de estimativa da evapotranspiração. Já em [10], foi usado o método para estimar a evapotranspiração para uso em um modelo para estimativa da umidade do solo. Segundo Zhu [25], o método SEBAL é adequado para estimativa da evapotranspiração para grandes regiões e, nesse trabalho, o método foi usado em todo o delta do Rio Amarelo, na China.

Alguns trabalhos fizeram uso de redes neurais para estimativa da evapotranspiração, como o feito por El-Shafie et al [5], que modificou uma rede neural e usou apenas dados das temperaturas máximas e mínimas para essa estimativa. Nesse trabalho, a rede neural foi treinada usando dados da evapotranspiração calculados a partir da equação de referência da FAO.

Por fim, o trabalho de Shiri et al [21] indica que um problema dos modelos gerados por mineração de dados sobre dados climáticos é que não podem ter sua performance inferida fora do local de treinamento/teste do modelo. Os autores propuseram um modelo de programação genética para estimar a evapotranspiração, que pode ser usado no local de treinamento ou fora dele, o que pode ser uma vantagem quando os dados não estiverem disponíveis em uma dada estação de medição.

Um aspecto comum a uma parte dos trabalhos relacionados nesta seção é a motivação usada para buscar outros métodos para estimar a evapotranspiração: conforme declarado em [4], o método de Penman-Monteith requer uma quantidade de variáveis que podem não estar disponíveis para o local de uso. Além disso, de acordo com a revisão feita em [15], pode não ser adequado para ser usado em áreas grandes, devido ao fato de não ser prático obter todos os dados necessários para o cálculo.

Na abordagem proposta no presente trabalho, a estimativa da evapotranspiração foi feita independentemente da disponibilidade de uma variável, garantindo flexibilidade no uso das variáveis. Além disso, através do uso de KDD, essa abordagem garante simplificação na estimativa da evapotranspiração, sem o uso de técnicas mais complexas, como as utilizadas pelo método SEBAL.

7. CONSIDERAÇÕES FINAIS

Pelo fato da equação de referência da FAO para estimativa da evapotranspiração ser dependente da disponibilidade dos valores de todas as variáveis da equação, buscou-se no presente trabalho uma forma de fazer essa estimativa independentemente de quais valores estejam disponíveis. Para alcançar esse objetivo, foi proposto o uso de descoberta de conhecimento em banco de dados, utilizando-se dados meteorológicos das seis estações de medição do INMET no Estado do Rio de Janeiro.

Para avaliação da solução proposta, foi feita uma comparação dos dados históricos da evapotranspiração potencial, disponíveis nos *datasets* utilizados, com os valores calculados pelos modelos aprendidos na execução do processo de KDD. A medida usada para avaliação para essa comparação foi o coeficiente de determinação (R^2), que varia entre 0 e 1, onde 1 indica uma correspondência de 100% entre os valores.

Nos *datasets* de cinco das seis estações usadas neste experimento, pelo menos um atributo estava sem os dados preenchidos. Uma alternativa seria preencher esses dados históricos através de modelos matemáticos tradicionais ou também usar mineração de dados para gerar um modelo de cálculo desses atributos. Optou-se, neste experimento, por não preencher os valores desses atributos vazios para analisar o comportamento do algoritmo nesses cenários.

Como resultado, para todas as estações, a medida de avaliação (R^2) esteve acima de 0.9, o que indica uma correspondência de mais de 90% entre o valor calculado e o valor histórico. Com este resultado, foi possível demonstrar que, com o uso de KDD, a evapotranspiração pode ser estimada com boa precisão, mesmo em cenários em que nem todos os valores estão disponíveis.

A abordagem utilizada neste trabalho é a sua principal contribuição, visto que o modelo é gerado a partir dos dados disponíveis e não o inverso. Dessa forma, a abordagem poderia ser utilizada para estimar a evapotranspiração em regiões que não dispõem de instrumentos de medição para todas as variáveis da equação de referência da FAO para estimativa da evapotranspiração.

Como outra contribuição, as equações geradas neste experimento podem ser utilizadas em trabalhos que necessitem de valores estimados da evapotranspiração como atributos de entrada para o cálculo de outra variável. Modelos como os gerados neste experimento também poderiam ser usados em cenários simulados, como os mantidos pelo IPCC.

Outra contribuição deste trabalho é que, para estações de medição com o foco no cálculo da evapotranspiração potencial, não seria necessário dispor de instrumentos de medição para todos os atributos meteorológicos, o que garantiria uma redução de custos na instalação de novas estações de medição. Pela característica do algoritmo de eliminar os atributos que contribuem pouco para o modelo, indicando apenas os que são realmente relevantes, têm-se informação de quais medidas realmente são necessárias para a estimativa da evapotranspiração.

Como trabalho futuro, a abordagem usada neste experimento poderia ser aplicada para estações de outras regiões do País. Uma questão a ser verificada é o desempenho do algoritmo escolhido para as estações do Estado do Rio de Janeiro em estações de outros estados do País, ou mesmo em outros países, onde as condições climáticas podem ser bem diferentes.

Outra possibilidade de trabalho futuro é o agrupamento de estações de acordo com algumas características. Esse agrupamento poderia ser usado para verificar se o modelo gerado para uma estação seria adequado para outra estação do mesmo grupo.

8. AGRADECIMENTO

Este trabalho foi parcialmente financiado pela FAPERJ (número do projeto: E-26/110.477/2014)

9. REFERÊNCIAS

- [1] Branco, P. M. (2014). Elementos que caracterizam o clima. Disponível <http://www.cprm.gov.br/publique/cgi/cgilua.exe/sys/start.htm?infoid=1267&sid=129> [acessado em 28-Novembro-2014].
- [2] Di Bello, R. C. (2005). Análise do Comportamento da Umidade do Solo no Modelo Chuva-Vazão Smap II-Versão com Suavização Hiperbólica Estudo de Caso: Região de Barreiras na Bacia do Rio Grande-BA (Dissertação de Mestrado, Universidade Federal do Rio de Janeiro).
- [3] Eagleson, P. S. (1994). The evolution of modern hydrology (from watershed to continent in 30 years). *Advances in water resources*, 17(1), 3-18.
- [4] El-Shafie, A., Najah, A., Alsulami, H. M., & Jahanbani, H. (2014). Optimized neural network prediction model for potential evapotranspiration utilizing ensemble procedure. *Water Resources Management*, 28(4), 947-967.
- [5] FAO (2014). Organização das Nações Unidas para Alimentação e Agricultura. Chapter 2 - fao penman-monteith equation. [Online; acessado em 10-Janeiro-2015].

- [6] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
- [7] Ferreira, A., & Meirelles, M. (2011). Implementação preliminar do modelo SEBAL para estimativa da evapotranspiração na Mesorregião do Sul Goiano. *Anais XV Simpósio Brasileiro de Sensoriamento Remoto*.
- [8] Haghverdi, A., Ghahraman, B., Leib, B. G., Pulido-Calvo, I., Kafi, M., Davary, K., & Ashorun, B. (2014). Deriving data mining and regression based water-salinity production functions for spring wheat (*Triticum aestivum*). *Computers and Electronics in Agriculture*, 101, 68-75.
- [9] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), 10-18.
- [10] Hendrickx, J. M., Pradhan, N. R., Hong, S. H., Ogden, F. L., Byrd, A. R., & Toll, D. (2009, May). Improvement of hydrologic model soil moisture predictions using SEBAL evapotranspiration estimates. In *SPIE Defense, Security, and Sensing* (pp. 730311-730311). International Society for Optics and Photonics.
- [11] Holman, D., Sridharan, M., Gowda, P., Porter, D., Marek, T., Howell, T., & Moorhead, J. (2013, August). Estimating reference evapotranspiration for irrigation management in the Texas high plains. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence* (pp. 2819-2825). AAAI Press.
- [12] Holman, D., Sridharan, M., Gowda, P., Porter, D., Marek, T., Howell, T., & Moorhead, J. (2014). Gaussian process models for reference ET estimation from alternative meteorological data sources. *Journal of Hydrology*, 517, 28-35.
- [13] INMET (2014). Instrumentos meteorológicos. Disponível em <http://www.inmet.gov.br/html/informacoes/sobre-meteorologia/instrumentos> (acessado em 28-Novembro-2014).
- [14] IPCC (2014). Intergovernmental panel on climate change, Disponível em <http://www.ipcc.ch> (acessado em 28-Novembro-2014).
- [15] Liou, Y. A., & Kar, S. K. (2014). Evapotranspiration estimation with remote sensing and various surface energy balance algorithms - A review. *Energies*, 7(5), 2821-2849.
- [16] Manesh, S. S., Ahani, H., & Rezaeian-Zadeh, M. (2014). ANN-based mapping of monthly reference crop evapotranspiration by using altitude, latitude and longitude data in Fars province, Iran. *Environment, development and sustainability*, 16(1), 103-122.
- [17] MMA (2015). Ciclo Hidrológico, Ministério do Meio Ambiente. Disponível em <http://www.mma.gov.br/agua/recursos-hidricos/aguas-subterraneas/ciclo-hidrologico>. (acessado em 06-março-2015)
- [18] Oliveira, M. D., & Carvalho, D. D. (1998). Estimativa da evapotranspiração de referência e da demanda suplementar de irrigação para o milho (*Zea mays* L.) em Seropédica e Campos, Estado do Rio de Janeiro. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 2(2), 132-135.
- [19] Quinlan, J. R. (1992, November). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence* (Vol. 92, pp. 343-348).
- [20] Recker, J. (2012). *Scientific research in information systems: a beginner's guide*. Springer Science & Business Media.
- [21] Shiri, J., Sadraddini, A. A., Nazemi, A. H., Kisi, O., Landaras, G., Fard, A. F., & Marti, P. (2014). Generalizability of Gene Expression Programming-based approaches for estimating daily reference evapotranspiration in coastal stations of Iran. *Journal of Hydrology*, 508, 1-11.
- [22] Tang, R., & Li, Z. L. (2015). Evaluation of two end-member-based models for regional land surface evapotranspiration estimation from MODIS data. *Agricultural and Forest Meteorology*, 202, 69-82.
- [23] Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes.
- [24] Zhou, X., Bi, S., Yang, Y., Tian, F., & Ren, D. (2014). Comparison of ET estimations by the three-temperature model, SEBAL model and eddy covariance observations. *Journal of Hydrology*, 519, 769-776.
- [25] Zhu, M., Hou, X., Lu, X., & Li, M. (2011, June). Spatial-temporal characters of evapotranspiration in the Yellow River Delta. In *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on* (pp. 409-412). IEEE.
- [26] Wikipedia (2015). Definição de R². Disponível em <http://pt.wikipedia.org/wiki/R%C2%B2> (acessado em 06-março-2015)

Metodologia de Geração de Plataforma de Inteligência de Negócios para Comissões de Avaliação de Universidade

Alternative Title: Methodology for Generation of Business Intelligence Platforms to University Evaluation Committees

Caio Eduardo Ribeiro

Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Brasil
31980-110
caioedurib@gmail.com

Anna Izabel J. Tostes Ribeiro

Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Brasil
31980-110
annatostes@gmail.com

RESUMO

As técnicas de Inteligência de Negócios visam extrair conhecimento estratégico dos grandes volumes de dados das organizações. Com a necessidade de se gerenciar e manusear em tempo hábil esses dados faz-se necessária uma plataforma de Inteligência de Negócios. Este artigo apresenta uma metodologia de geração de plataformas para automação do processo organizacional de Comissões de Avaliação de universidade, que automatizam parte do processo e auxiliam na geração de conhecimento estratégico. A metodologia proposta abrange desde a especificação até a avaliação do protótipo da plataforma, feita segundo a norma ISO/IEC 9126. Foi criado um estudo de caso em uma Comissão de Avaliação de uma universidade, e plataforma desenvolvida obteve uma média geral de 88,9% de aprovação em sua avaliação, sendo considerada apta para implantação.

Palavras-chave

Inteligência de Negócios, Automação, Projeto de Software

ABSTRACT

Business Intelligence techniques aim to extract strategic knowledge from great data volumes in organizations. With the need to manage and handle timely this data, a Business Intelligence Platform becomes necessary. This paper presents a methodology for generation of platforms for automation of organizational process from University Evaluation Committees, which automate part of the process and support strategic knowledge generation. The proposed methodology comprehends from specification to evaluation of the platform prototype, made according to ISO/IEC 9126 standard. A case study in an University Evaluation Committee was made, and the developed platform had an 88.9% approval rate, considered ready for deployment.

Categories and Subject Descriptors

H.3.5 [Online Information Services]: Web-based services

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.
Copyright SBC 2015.

J.1 [Administrative Data Processing]: Education

General Terms

Measurement, Design, Standardization.

Keywords

Business Intelligence, Automation, Software Project

1. INTRODUÇÃO

As plataformas de Inteligência de Negócios (*Business Intelligence* – BI) combinam coleta e armazenamento de dados, gerenciamento de conhecimento e ferramentas de análise para apresentar informações internas e competitivas aos tomadores de decisão [11]. O grande volume de dados gerados pelas organizações torna difícil seu armazenamento e manuseio. Está implícita nessa definição a ideia de que plataformas de BI devem prover essas informações de forma segura e em tempo hábil, enquanto ainda são válidas para auxiliar na tomada de decisão [7].

A Comissão Própria de Avaliação (CPA) de uma universidade, tornada obrigatória pela Lei [15], é responsável por fazer todas as avaliações no âmbito da universidade, e encaminhá-las ao Ministério da Educação, com o objetivo de manter a qualidade das instituições de ensino superior. As informações geradas pela análise dos resultados das avaliações aplicadas pela CPA devem ser entregues em tempo hábil para que possam ser úteis no auxílio à tomada de decisões administrativas da instituição. A falta de automação de alguns processos internos que demandam muito tempo dos membros da Comissão pode atrasar a entrega dos resultados.

Este artigo objetiva detalhar a metodologia de geração, e avaliação do protótipo de um sistema de BI desenvolvido na plataforma Sharepoint, com o objetivo de gerenciar e integrar os processos da Avaliação Institucional feita pela CPA da Pontifícia Universidade Católica de Minas Gerais (PUC Minas). Os resultados esperados são a automatização de parte do processo das avaliações realizadas, e a governança dos dados eletrônicos da organização. A plataforma desenvolvida foi avaliada segundo a norma ISO/IEC 9126 de avaliação de qualidade de software, e considerada apta para implantação.

Este artigo é apresentado em sete seções. Os conceitos de Inteligência de Negócios necessários para a compreensão do trabalho estão na seção 2. A seção 3 indica trabalhos relacionados a este. A seção 4 contém uma descrição da metodologia de geração de software proposta, e a seção 5 apresenta um

detalhamento do estudo de caso realizado para a CPA da PUC Minas. Os resultados da avaliação do sistema se encontram na seção 6. Finalmente, a seção 7 contém as conclusões e direções futuras para o trabalho.

2. INTELIGÊNCIA DE NEGÓCIOS

Inteligência de Negócios é uma evolução natural de uma série de sistemas anteriores, criados para dar suporte à tomada de decisões [7]. Esses sistemas podem ser divididos em seis categorias que funcionam de forma cíclica. Ou seja, os sistemas de uma categoria fornecem as entradas para os da categoria seguinte, até que uma nova iteração se inicie, conforme ilustrado na Figura 1.

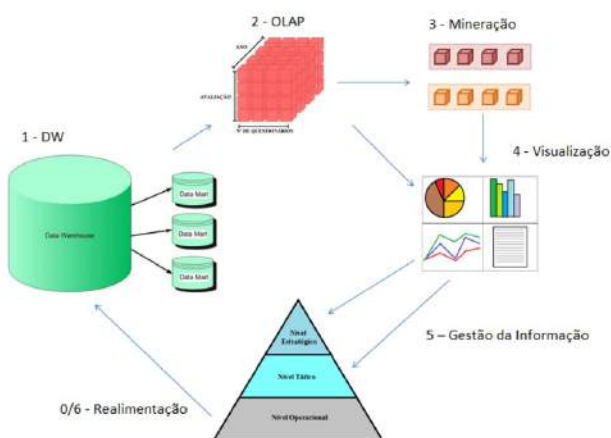


Figura 1. O Ciclo da Inteligência de Negócios

Um *Data Warehouse* (DW) é um grande Banco de Dados que contém os dados referentes às operações diárias da organização. Os DW são direcionados principalmente para aplicações de apoio à decisão, sendo otimizados para a recuperação de dados, e não para o processamento de transações como os Bancos de Dados (BDs) tradicionais [4]. Os DW possuem foco de auxílio à decisão, e servem às aplicações *On-Line Analytical Processing* (OLAP), que otimizam a recuperação de dados multidimensionais para auxiliar análises estatísticas de diversas visões da organização.

Mineração de Dados (*Data Mining*) é o processo de identificar e interpretar padrões nos dados para se resolver um problema específico na organização. As ferramentas de Mineração utilizam técnicas de inteligência artificial e estatística avançada para descobrir fatos sobre os dados de forma automatizada [11]. As etapas do processo de *Knowledge Discovery in Databases* (KDD) proposto por [13] são Seleção, Pré-processamento, Transformação, Mineração de Dados e Avaliação, ou seja, a realização de mineração é necessária para gerar conhecimento através de uma Base de Dados.

Visualização é o nome do processo de representação de dados através de imagens gráficas. Tecnologias de visualização dão uma visão global de dados complexos, explorando a capacidade de sintetização da visão humana para criar imagens que representem as informações que devem ser transmitidas [7].

Estando as informações devidamente representadas, cabe às tecnologias de Gestão da Informação identificar o conhecimento que elas geram, e distribuí-lo adequadamente [10].

Os gráficos, painéis e relatórios dinâmicos que o processo realizado gerou vão chegar às pessoas que tomarão decisões, auxiliando nas mesmas. As decisões são tomadas nos níveis Tático e Estratégico da organização, e propagadas para os níveis inferiores, podendo modificar os processos do nível Operacional. Novos dados são gerados pelos processos de rotina da organização, e estes são carregados no DW, reiniciando o ciclo.

3. TRABALHOS RELACIONADOS

Esta seção apresenta os principais trabalhos da literatura relacionados a este trabalho. O grupo Gartner, criador do termo *Business Intelligence*, avalia as plataformas de BI no artigo [8] e as classifica em categorias de acordo com a completude da visão e a habilidade de execução da plataforma. Na avaliação de 2013 a ferramenta Microsoft Sharepoint foi considerada a plataforma com maior habilidade de execução, e posicionada na categoria das plataformas líderes, pois sua completude de visão também é uma das melhores.

A arquitetura de um sistema de BI baseado em serviços web é apresentada em [9]. Os autores defendem que, no cenário empresarial atual, a obtenção de informações de suporte a decisão precisa ser ágil, e os sistemas precisam responder rapidamente às constantes mudanças dos requisitos do negócio e do ambiente. Um sistema baseado em serviços web possui essas e outras características desejadas, como a capacidade de coordenar a operação de várias plataformas ao mesmo tempo, boa escalabilidade, facilidade de implantação e adaptabilidade adequada para atender as mudanças no negócio. A arquitetura proposta no artigo é baseada na arquitetura de sistemas tradicionais de BI, e dividida em três camadas: uma camada Expressional, que contém a interface do sistema com os usuários; a camada Funcional onde os dados são tratados por funções de BI para serem apresentados; e a camada de Dados onde os servidores, estão localizados.

O artigo [6] descreve o desenvolvimento de uma solução de BI centrada em um DW, que utiliza integração de dados e OLAP para permitir, através de relatórios e gráficos, a visualização de dados financeiros e estratégicos de uma empresa em tempo real, dando apoio ao processo de tomada de decisão. Os autores descrevem as etapas do desenvolvimento do software e apontam os fatores principais para que um projeto de BI seja bem sucedido: definição clara dos objetivos, envolvimento das partes interessadas no projeto, estabelecimento de cronogramas e orçamentos e administração de expectativas. A plataforma de arquitetura escolhida para desenvolver o sistema foi o Microsoft Sharepoint, pois os relatórios são acessados via web, o que permite o acesso à informação em tempo hábil. Os autores constataram que o uso da interface web possui algumas limitações, como a manipulação limitada das dimensões de dados (os relatórios disponíveis são pré-formatados), e isso tornou necessário o uso de ferramentas específicas para a realização de algumas análises.

O estudo destes e outros trabalhos contribuiu para a formação da estratégia adotada na metodologia proposto, uma ferramenta web de automação de parte do processo operacional de uma CPA da PUC Minas, implantada na plataforma Sharepoint 2010. O objetivo do sistema é a otimização do processo interno e a realização da governança dos dados eletrônicos da organização. A metodologia desenvolvida tem foco nas possibilidades de um

sistema web, principalmente na modularidade que a disponibilidade das informações permite.

4. METODOLOGIA

Esta seção detalha a metodologia de geração de software criada com o intuito de guiar o projeto, implantação e avaliação de sistemas de BI para Comissões de Avaliação de universidade. As ferramentas web precisam ter equilíbrio entre dinamismo e robustez, aptas a lidar com as constantes mudanças no negócio e otimizar o processo operacional da organização.

Para se aproveitar da flexibilidade e paralelismo possibilitados por uma plataforma web, é necessário que o processo não seja “engessado”, e algumas de suas etapas possam ser realizadas ao mesmo tempo por pessoas ou equipes diferentes. No caso de uma avaliação de universidade, existem muitos aspectos a serem observados, e se eles forem examinados um a um, a avaliação se torna lenta e pesada. Por isso, este modelo propõe o conceito de Dimensões.

As Dimensões de uma avaliação são os aspectos diferentes da universidade que precisam ser verificados para garantir a qualidade do ambiente de ensino. O Ministério da Educação criou uma série de indicadores de qualidade que devem guiar a concepção das avaliações realizadas pela CPA, e através destes, é possível dividir a avaliação em uma série de Dimensões que podem ser consideradas individualmente. Isso permite que, por exemplo, uma equipe trabalhe na avaliação da dimensão Infraestrutura da Universidade ao mesmo tempo que outra trabalhe na avaliação da dimensão Pesquisa e Extensão.

A principal contribuição desta metodologia é o processo ser modularizado de forma a permitir que grande parte da descoberta de conhecimento seja realizada em paralelo, ou seja, partes do objetivo principal do processo serem produzidos ao mesmo tempo por equipes diferentes. Como a plataforma web integra o processo operacional da organização em um único sistema online, as informações estão disponíveis assim que são produzidas, e os fragmentos do produto final podem ser produzidos simultaneamente para posteriormente serem integrados, graças ao conceito de Dimensões. Além disso, a plataforma embarca as funções do KDD, automatizando-as na medida do possível, podendo no melhor caso fazer todas as operações necessárias *on-line*.

A Figura 2 mostra como o processo de KDD se torna mais dinâmico quando embarcado em uma plataforma de BI. As etapas de seleção, pré-processamento, transformação e mineração dos dados são realizadas pelas funções internas da plataforma, e as técnicas de visualização de BI facilitam sua interpretação para geração de conhecimento.



Figura 2. KDD na plataforma web de BI (adaptada de [13])

A metodologia criada é uma adaptação das diretrizes clássicas de projeto e desenvolvimento de software, com foco nos objetivos de sistemas web de automação e BI. Ela foi dividida em cinco fases interligadas, que se sobrepõem em partes da execução da metodologia, e sete diretrizes de projeto que se entrelaçam com essas fases.

4.1 Fases da Metodologia

A primeira fase é a de definição do Processo. Pode ser necessário um mapeamento do processo interno da organização, caso não haja documentação do mesmo, para determinar quais seus pontos críticos e como o software irá abordá-los. Se a organização já tiver um processo maduro e bem definido, a plataforma tem um desenvolvimento mais simples, mas este não é um caso comum especialmente no cenário brasileiro.

É de responsabilidade da equipe de desenvolvimento coletar informações sobre o processo operacional da organização, para traçar estratégias de como melhorá-lo com a plataforma de BI, e essa definição prévia é importante para que o produto gerado seja o que o cliente espera. Durante essa etapa, as avaliações devem ser divididas em Dimensões de acordo com os Indicadores do MEC abordados por elas.

Em seguida, é feito o mapeamento dos Dados. Nessa etapa é definido como os dados do sistema serão adquiridos e transformados. Os softwares de organizações frequentemente são responsáveis por prover acesso a dados vindos de diversas fontes, e é importante que haja um controle desse acesso, que também é definido nesta etapa. É importante ressaltar que, em um sistema de BI, é comum que os dados sejam gerados com alta frequência e grande volume, e o sistema deve estar preparado para lidar com isso.

A terceira fase é a do Projeto do software, que deve determinar exatamente o seu escopo e objetivos. O projeto é usado como guia durante as etapas seguintes para garantir que a equipe de desenvolvimento e os clientes estejam alinhados quanto ao produto final do projeto. Cada solução do sistema deve ser projetada individualmente, com seus dados de entrada e saída, e as operações necessárias para alcançar seus objetivos. Com o projeto modular, é possível o desenvolvimento paralelo das soluções do sistema pela equipe. Esta etapa também envolve a definição de como o sistema irá se apresentar para o usuário: seu *layout* (preferivelmente com ícones e botões, e no formato de uma página web como as que os usuários estão habituados). Finalmente, deve ser determinado de que maneira o usuário será treinado a utilizar o sistema e as formas de ajuda que estarão disponíveis.

Com a finalização do projeto, entra a fase de Desenvolvimento. Todos os sites da plataforma web são desenvolvidos, e as soluções nativas do Sharepoint e as customizadas para o sistema são implementadas de acordo com o projeto realizado. Entregas parciais são recomendadas, porque elas permitem que o projeto amadureça antes do produto final ser entregue. O principal objetivo de um sistema de BI é a geração de conhecimento estratégico em tempo hábil, e essa deve ser a principal preocupação da equipe de desenvolvimento durante todo o processo, especialmente na otimização de suas soluções.

Neste tipo de projeto, é comum que hajam alterações nos requisitos do sistema durante a etapa de desenvolvimento, por

mais que estas sejam evitadas nas etapas anteriores. A equipe de desenvolvimento deve estar preparada para fazer controle de alterações e definir como elas vão impactar no projeto como um todo.

Finalizando o processo, temos a etapa de Avaliação. Para determinar se o sistema atingiu os objetivos definidos em seu projeto, é feita uma avaliação quantitativa que reflete a qualidade do software em números que podem ser comparados com avaliações de versões futuras, e guiar a equipe de desenvolvimento nos próximos projetos para o sistema. A avaliação é feita seguindo a norma ISO/IEC 9126, e pode ser aplicada em forma de questionário ou entrevista, dependendo do projeto. As cinco etapas da metodologia são ilustradas pela Figura 3.

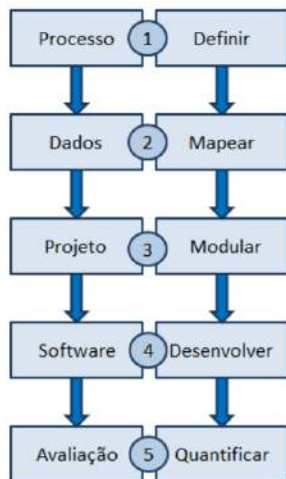


Figura 3. Etapas da metodologia de geração de software

4.2 Diretrizes para o projeto da plataforma

Antes de se iniciar o desenvolvimento de uma plataforma de BI, é muito importante que ela seja projetada adequadamente. Os *stakeholders* (pessoas envolvidas com o projeto, principalmente os clientes) e os desenvolvedores precisam entrar em acordo prévio sobre todos os detalhes do produto a ser entregue. Para o desenvolvimento de uma estratégia para o projeto de uma plataforma feita em Microsoft Sharepoint, são apresentados em [3], sete diretrizes a partir das quais foi definida estratégia guia para as cinco fases da metodologia:

1. **Propósito (Purpose):** Identificação dos objetivos da plataforma, que precisam estar bem definidos antes do início da implementação. Abordado na primeira e na terceira fases.
2. **Controle (Governance):** Compreende uma documentação dos processos, padrões, papéis, responsabilidades e as políticas que são usadas na organização. Em suma, todas as regras e características do negócio. Abordado na primeira e segunda fases.
3. **Pessoas e Objetivos (People and Objectives):** Identificação das audiências que possuem participação no ambiente da plataforma. É importante que se compreenda quem são essas pessoas, e quais são as

tarefas que elas precisam desempenhar. Abordada na primeira fase.

4. **Requisitos e Análise (Requirements and Analysis):** Através de entrevistas com os *stakeholders* do sistema, reuniões das equipes de desenvolvimento com os clientes, entrevistas com usuários finais, entre outros, os requisitos do sistema são formalmente coletados e documentados. Abordada na terceira fase.
5. **Arquitetura da Informação (Information Architecture):** Define as abordagens de duas áreas-chave: Organização da Informação e Acesso à Informação. Importante para a realização da governança dos dados da organização. Abordada nas segunda e terceira fases.
6. **Tecnologia (Technology):** As considerações de tecnologia incluem instalação, configuração e manutenção de ambos o *hardware* e o software necessários para o uso da plataforma, bem como a integração do ambiente Sharepoint com os sistemas da organização. Abordada na terceira e na quarta fases.
7. **Manutenção e Melhoria (Maintenance and Enhancement):** Compreende métricas de avaliação de desempenho da aplicação, que avaliam tanto a satisfação do usuário final quanto o *benchmark* da plataforma. Para lidar com as mudanças que se mostrarem necessárias, enquanto o negócio se modifica e evolui, a plataforma precisa ter suas funcionalidades constantemente avaliadas, permitindo que os desenvolvedores tomem decisões relativas às mudanças da aplicação. Abordada na última fase.

O planejamento da plataforma deve ser uma tarefa minuciosa, para que todos os detalhes do processo estejam bem definidos previamente. A Engenharia de Software recomenda não se poupar tempo e recursos nessa etapa, porque gerenciar mudanças de projeto durante seu desenvolvimento pode prejudicar bastante o cronograma e até inviabilizar o projeto. Isso é ainda mais crítico em uma plataforma que abarca tantas partes do processo de uma organização, desde a gestão dos seus dados e pessoas até a geração de produtos e conhecimento. Deve-se atentar para as áreas mais críticas e que possuem necessidades mais abrangentes de informações gerenciais [Barbieri 2001].

As primeiras reuniões entre a equipe de desenvolvimento e os *stakeholders* são para definir o Propósito da plataforma, ou seja, seus objetivos perante a organização e suas atividades. Definidos os resultados esperados do projeto, as reuniões seguintes devem explicitar as características de Controle e as Pessoas e Objetivos relevantes. É recomendado que sejam criados grupos de usuários, para garantir a seus membros o acesso a funcionalidades do sistema, e cada usuário deve ser alocado em um ou mais grupos, de acordo com as tarefas que deve desempenhar.

A diretriz seguinte é a de Requisitos e Análise na qual, através de entrevistas com os *stakeholders* do sistema, reuniões das equipes de desenvolvimento com os clientes, entrevistas com usuários finais, entre outros, os requisitos dos softwares da plataforma são formalmente coletados e documentados. Os requisitos devem ser classificados em três categorias, de acordo com sua relevância. São considerados Essenciais os requisitos imprescindíveis para que o software funcione, Importantes aqueles sem os quais o software funciona, mas de forma não

satisfatória, e Desejáveis os que não comprometem as funcionalidades básicas do software, e podem ser deixados para versões futuras do sistema caso não haja tempo hábil para implementá-los na versão que está sendo especificada.

Para a descrição da Arquitetura da Informação, são definidos os dados que a plataforma precisará gerenciar, e como será sua organização. Antes da implementação iniciar, a equipe deve ter uma ideia clara de como irá obter e administrar esses dados, e escolher uma arquitetura que aplique corretamente as regras do negócio, além de definir as ferramentas de gestão de dados que serão utilizadas.

Nesta metodologia, foi definido que seria utilizado o padrão de projeto MVC (padrão *Model-View-Controller*, amplamente adaptado para uso em aplicações Web [14]) para o desenvolvimento das Web Parts Customizadas, softwares desenvolvidos especialmente para cumprir uma função do processo, que são embarcados na plataforma pelo Sharepoint. Durante a descrição da Arquitetura da Informação, também são especificadas as ferramentas a serem utilizadas e a forma de comunicação entre a plataforma e suas fontes de dados. Essa parte da diretriz e a diretriz de Tecnologia são realizadas em conjunto, pois a equipe de Tecnologia da Informação da universidade é responsável por considerar as necessidades de hardware e software que as decisões de arquitetura implicam.

5. Estudo de caso

Para testar a metodologia proposta, foi feito um estudo de caso com o desenvolvimento de uma plataforma de BI para a CPA da PUC Minas. A etapa de avaliação foi descrita durante a apresentação dos resultados da avaliação dessa plataforma, na seção seguinte. Não serão apresentadas todas as etapas do processo, pois para o escopo deste trabalho é necessário abordar apenas as essenciais para a compreensão do sistema desenvolvido e como a metodologia auxiliou em seu projeto, implantação e avaliação.

5.1 Melhoria no processo de avaliação

A plataforma desenvolvida é um protótipo que inicialmente tem como foco automatizar os processos operacionais realizados na CPA da PUC Minas, porém foi planejada de maneira escalável, permitindo que em trabalhos futuros sejam acrescentadas novas funcionalidades.

Foi definido que o sistema deveria automatizar a Avaliação Institucional realizada pela organização, pois esse é o processo mais custoso realizado na CPA da PUC Minas. A Figura 4 mostra as principais etapas do processo de avaliação abordado, antes de ser modificado para o sistema. As cores diferentes indicam sistemas dos quais o processo dependia para seguir adiante, o que muitas vezes causava atrasos devido à mão de obra especializada necessária, e necessidade de colaboração de equipes envolvidas em outros projetos. Grande parte do processo era realizado através do trabalho manual dos funcionários da CPA, e não havia nenhum sistema que gerenciasse os documentos eletrônicos da organização, causando problemas de versionamento e disponibilidade dos dados.

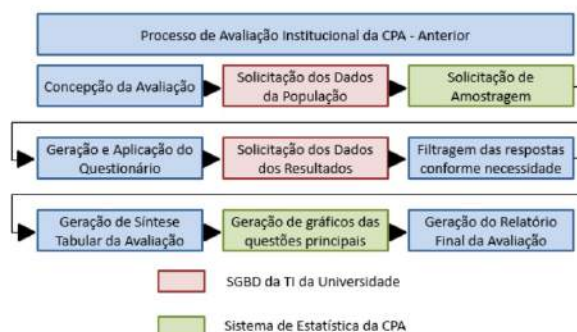


Figura 4. Processo de Avaliação anterior

Nessa etapa, foram identificadas as fontes de dados e como deveria ser a comunicação entre a plataforma e os demais sistemas para que a plataforma mantivesse o controle sobre o processo. Uma decisão importante foi aumentar a granularidade dos relatórios da avaliação, o principal produto do processo, proposta trazida por este trabalho. Com o objetivo de permitir que partes do relatório fossem geradas em paralelo por usuários diferentes, foram definidas as Dimensões da avaliação em questão, e esse conceito foi usado na implementação das soluções desenvolvidas.

O processo modificado para a Avaliação Institucional é exibido na Figura 5. Foram desenvolvidas Web Parts na plataforma para gerar automaticamente a amostra, os gráficos, as Sínteses Tabulares e o Relatório Final (reunindo dados das Dimensões, para edição dos especialistas da CPA), e todos os dados necessários para a realização das operações da organização foram negociados com seus responsáveis, e centralizados em uma Base de Dados local, sob gestão da TI da universidade. Dessa forma, o processo passou a não depender de outros sistemas. Além disso, a plataforma deveria fazer a gestão dos documentos eletrônicos da organização, incluindo o controle de acesso aos mesmos e seu versionamento. O processo modificado é reduzido e automatizado em grande parte, tornando a operação da organização mais ágil e eficiente, e permitindo mais atenção à qualidade do produto.



Figura 5. Processo de Avaliação Modificado

5.2 Plataforma desenvolvida

A navegação do SIAV se baseia principalmente no menu de *links* encontrado à esquerda da tela de todas as páginas, chamado *Quick Launch*, que é sensível ao contexto, e exibe os *links* mais relevantes para o usuário de acordo com onde ele se encontra no sistema. O conteúdo das páginas da aplicação é facilmente

alterável pelos usuários com permissão de edição de página, podendo ser adaptado às mudanças dos processos sem a necessidade do desenvolvimento de uma nova versão. Isso torna a ferramenta adaptável e flexível, necessidade inerente a sistemas com grande teor de modificações, e diminui a necessidade de manutenções custosas que interrompem o processo operacional.

Todas as soluções nativas utilizadas no SIAV se baseiam nas funcionalidades de Listas do Sharepoint, definidas como uma coleção de informações compartilhadas entre membros de uma equipe. As Listas predefinidas do Sharepoint incluem listas de eventos, calendários, quadros de discussão, além da possibilidade de conectar as listas a fontes de dados externas e de criar listas customizadas.

Um tipo especial de lista é a Biblioteca de Documentos, otimizada para armazenar documentos, planilhas, slides ou outros tipos de arquivos, e foi esse tipo de lista o escolhido para facilitar a gestão eletrônica dos documentos da CPA. Usando essa lista, os usuários têm garantia de que todos estão trabalhando com os mesmos arquivos, podem ver as alterações realizadas nos mesmos, e ter a segurança de apenas uma edição estar sendo feita naquele momento (funcionalidade de *check-in* e *check-out*). Através dessas listas, o sistema realiza a governança de dados da organização, centralizando-os em um só servidor e permitindo seu versionamento e controle de alteração.

Existem partes do processo da avaliação que não podem ser realizadas pelas soluções nativas do Sharepoint. Para essas partes, foram desenvolvidas soluções customizadas, programas com funções que vão tornar o processo da CPA mais rápido e eficiente, com alto grau de confiabilidade, visto que os programas são menos suscetíveis a erros do que o trabalho humano.

As Web Parts customizadas foram desenvolvidas nas linguagens ASP.NET e C#, na IDE Visual Studio 2010, que é a recomendada pela Microsoft para a construção de soluções de Sharepoint 2010 [12]. Para realizar as consultas ao BD necessárias, foi utilizada a biblioteca gratuita SqlConnection, que cria uma sessão de conexão com o Banco para aquela aplicação, executa *queries* e retorna seus resultados dentro do contexto da conexão estabelecida. Para a geração dos documentos do Word 2010 e planilhas do Excel 2010 criados pelas Web Parts, foi utilizada a biblioteca gratuita Open XML SDK, que permite a escrita de um arquivo XML (*eXtensible Markup Language*, o formato padrão dos arquivos do Microsoft Office 2010) sem a necessidade de utilizar os programas do Office para criar os arquivos. Finalmente, para a geração de imagens JPEG (gráficos gerados automaticamente) foi utilizada a biblioteca gratuita Interop.Excel, da Microsoft, que inicia uma instância do Excel na máquina do servidor, e parametriza a criação e exportação da imagem do gráfico.

Na CPA da PUC Minas, uma das primeiras etapas do processo de avaliação por questionário é a seleção do conjunto de respondentes, ou seja, dos indivíduos para os quais o questionário será aplicado. A esse conjunto é dado o nome de amostra, e sua seleção é sempre aleatória dentre um grupo predefinido de indivíduos, chamado conjunto universo. A Web Part de Geração de Amostras permite que o usuário selecione um ou mais conjuntos-universo (conjunto de registros a ser considerado o total para aquela categoria), e o número de respondentes desejado para cada amostra a ser gerada, e faz a seleção aleatória para o usuário, criando uma planilha Excel com os dados dos registros selecionados.

Para facilitar a visualização dos dados obtidos pela aplicação de questionários, os relatórios da CPA apresentam esses dados na forma de gráficos, que são analisados pelos especialistas da Comissão, compondo o relatório de avaliação. A geração desses gráficos é realizada por uma Web Part que permite ao usuário selecionar uma ou um grupo de questões de algum questionário aplicado, e em seguida informe as configurações desejadas para o gráfico a ser gerado. Para cada gráfico é adicionada automaticamente uma imagem na Lista de Imagens, com uma numeração correspondente à Dimensão do relatório a qual ela pertence informada no nome da imagem. Além da imagem, também é gerada uma planilha-fonte com os dados que foram usados na construção do gráfico, para conferência, e esta é adicionada a uma lista de planilhas-fonte, na mesma página.

Além do relatório de avaliação, a CPA gera documentos de Síntese Tabular para facilitar a análise dos resultados obtidos pelos interessados. O documento de Síntese possui uma tabela para cada questão selecionada, apresentando os resultados obtidos para aquela questão de forma sucinta. Para gerar a Síntese, o usuário deve selecionar o conjunto de questões e escolher o nome do arquivo, e uma Web Part customizada criada para gerar esses documentos cria e adiciona a Síntese à lista destas, presente na mesma página.

O produto principal da Avaliação Institucional é seu relatório final, que contém as análises da CPA acerca dos resultados obtidos na aplicação do questionário. Esse relatório é formado por várias partes, os chamados Documentos de Dimensão, que ficam armazenados na Biblioteca da página de Dimensões do site. Cada item dessa lista é um documento que será integrado no relatório final. Além dos documentos de Dimensão, o relatório também contém os gráficos e análises dos resultados das questões dos questionários aplicados naquela avaliação. Para facilitar o trabalho de construção do relatório, uma Web Part de geração de relatório acessa as listas de Documentos de Dimensão e Imagens do site, e monta um relatório parcial, reunindo esses fragmentos.

Em resumo, o SIAV realiza a governança dos dados da organização, permite a colaboração *on-line* dos usuários, realiza o controle de acesso às funcionalidades automaticamente, e automatiza partes importantes do processo. Como é um sistema web, a flexibilidade e possibilidade de expansão do SIAV são adequadas para o cenário de mudanças constantes de uma organização cujo processo está em constante evolução.

6. AVALIAÇÃO QUANTITATIVA DA PLATAFORMA

Uma avaliação quantitativa deve permitir uma noção exata do quanto o software cumpre as expectativas dos clientes. O SIAV foi avaliado de acordo com a norma ISO/IEC 9126 [1], criada com o intuito de servir como uma padronização da avaliação da qualidade de produtos de software, que estabelece uma avaliação quantitativa que pode ser repetida a cada nova versão da plataforma.

A norma ISO/IEC 9126 tem foco na qualidade do produto de software, propondo Atributos de Qualidade, distribuídos em seis características principais, cada uma delas divididas em sub-características que serão avaliadas, exibidas na Figura 6. O método escolhido para a avaliação do SIAV foi a

entrevista, devido ao número reduzido de usuários (dez funcionários da organização, que iriam utilizar o sistema).

ISO/IEC 9126	Funcionalidade	Adequação	Segurança
		Acurácia	Conformidade
		Interoperabilidade	
	Confiabilidade	Maturidade	Recuperabilidade
		Tolerância a Falhas	Conformidade
	Usabilidade	Inteligibilidade	Atratividade
		Aprensibilidade	Conformidade
		Operabilidade	
	Portabilidade	Adaptabilidade	Substituição
		Facilidade de Instalação	Conformidade
		Coexistência	
	Manutenibilidade	Analísabilidade	Testabilidade
		Modificabilidade	Conformidade
		Estabilidade	
	Eficiência	Consumo de Recursos	Conformidade
		Consumo de Tempo	

Figura 6. Características da ISO/IEC 9126

Finalizando as etapas da metodologia de geração de software proposta, foram feitas entrevistas com os funcionários da CPA nas quais cada entrevistado deveria dar uma nota para a plataforma avaliada, para cada sub-característica das características avaliadas na norma ISO/IEC 9126. O foco do projeto da plataforma foi na sua funcionalidade e usabilidade, características que determinam o índice de adesão da plataforma, porque o sistema está em sua primeira versão, e ainda não havia sido implantado. Para as características mais importantes, foi dado um peso maior, de forma que as médias obtidas nestas duas características valessem duas vezes mais no cálculo da média geral obtida pelo sistema. Se essas notas fossem decididas pelo avaliador, analisando as respostas dos usuários de perguntas sobre o sistema, os resultados poderiam se tornar tendenciosos, portanto a entrevista reflete melhor a opinião do usuário se a nota for decidida no momento da entrevista, em conjunto com ele [5] (naturalmente, o papel do entrevistador é apenas tirar dúvidas a respeito do conceito a ser avaliado, sem influenciar de maneira alguma na resposta do entrevistado).

Foram entrevistados 10 funcionários da CPA, e a média aritmética simples das notas de cada entrevistado foi considerada a nota da plataforma para a sub-característica em questão. A média aritmética simples das notas das sub-características foi considerada a nota final da plataforma para característica, e a média geral é uma média ponderada das notas de todas as características, considerando um peso de dois para as características de Funcionalidade e Usabilidade.

Antes da realização das entrevistas, a plataforma foi apresentada para o entrevistado, e suas funcionalidades foram explicadas detalhadamente, além dos objetivos estabelecidos no projeto. Em seguida, foi explicado ao entrevistado cada sub-característica a ser avaliada, e ele poderia escolher uma nota entre 0 e 10, ou se abster de dar sua opinião para aquela sub-característica, caso não achasse que haviam informações suficientes para formá-la. Não houveram limitações de tempo para a resposta das perguntas. Os entrevistados tiraram suas dúvidas sobre a plataforma no momento da entrevista, e deram sugestões de melhorias que foram discutidas em reuniões com os *stakeholders* e, caso aprovadas, implementadas em seguida.

Foi estabelecido um objetivo de 80% de aprovação para que a plataforma fosse considerada totalmente aprovada. O gráfico da Figura 7 mostra os resultados obtidos pela plataforma na avaliação de cada característica, e a média geral final. O SIAV foi aprovado totalmente em sua avaliação geral, e a única característica que não atingiu a meta foi a de Confiabilidade. Isso ocorreu porque a sub-característica de Tolerância a Falhas, que avalia a reação da plataforma em situações adversas, recebeu notas baixas, pois ainda não haviam sido implementadas funções de comunicação eficientes, que informassem ao usuário que ações tomar em caso de falha. Aumentar a confiabilidade do sistema deve ser, portanto, um dos focos no projeto das próximas versões.

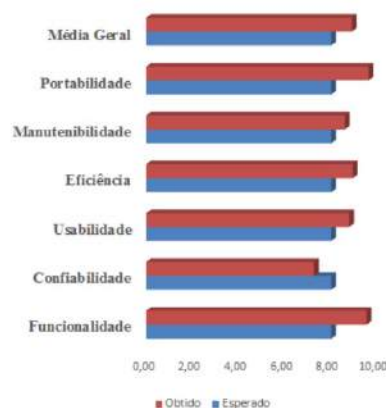


Figura 7. Resultados Obtidos na Avaliação do Sistema

7. CONCLUSÕES

Este trabalho apresentou uma metodologia para guiar o projeto e desenvolvimento de plataformas de BI que automatizam parte do processo operacional de Comissões de Avaliação de universidade. Foram descritos os detalhes de estruturação da plataforma e as tecnologias utilizadas, de forma que ela possa ser replicada no projeto de plataformas com características semelhantes. A plataforma desenvolvida para testar a metodologia foi avaliada segundo a norma ISO/IEC 9126, e obteve uma pontuação média de 88,9%, sendo considerada aprovada.

O uso do Microsoft Sharepoint para desenvolvimento de plataformas web de colaboração e governança de dados é endossado porque, apesar do alto custo financeiro e da necessidade de técnicos treinados para administrar a aplicação, a ferramenta traz flexibilidade, disponibilidade e grande capacidade de execução para o projeto. As aplicações podem ser aprimoradas e passar por manutenção sem ter seu funcionamento interrompido, a maior parte do tempo. Os documentos da organização são centralizados, aumentando sua disponibilidade e diminuindo o custo de armazenamento e gestão, sem perder o controle de acesso e versionamento dos mesmos.

Durante a modelagem do processo operacional da organização, foram identificadas as funções críticas e como torná-las mais eficientes, e foi decidido que o processo da avaliação deveria ser dividido em Dimensões e indicadores, o que tornou o processo mais dinâmico e permitiu a realização paralela de algumas de suas etapas. A plataforma desenvolvida automatiza parte dos processos da organização, além de realizar a gestão eletrônica de seus documentos. A estrutura do SIAV é adaptável

para qualquer avaliação baseada em questionário, o que torna o sistema expansível para os outros processos da CPA além dos abordados, e faz com que a metodologia apresentada possa ser adotada sem muitas alterações por outras organizações que utilizam essa forma de avaliação.

As direções futuras deste trabalho incluem a implantação de um *Data Warehouse* da CPA, para que as avaliações possam apresentar a progressão de vários aspectos da universidade ao longo do tempo. O sistema também será adaptado para incluir módulos para as outras avaliações realizadas pela organização, e um módulo de gestão de processos internos.

8. REFERÊNCIAS

- [1] Abran, A. 2010. *Software Metrics and Software Metrology*. John Wiley & Sons Inc.
- [2] Barbieri, C. 2001. *BI-Business Intelligence: modelagem e tecnologia*. Axxel Books
- [3] Carr, J. 2011. *Case study: Developing a Sharepoint 2010 strategy. . . or how setting it up and getting it out there is not a strategy*. Bulletin of the American Society for Information Science and Technology, v. 37, p. 26–28.
- [4] Chaudhuri, S.; Dayal, U. 1997. *An overview of Data Warehousing and OLAP technology*. SIGMOD Rec., ACM, New York, NY, USA, v. 26, n. 1, p. 65–74.
- [5] Lakatos, E., Marconi, M. 2010. *Fundamentos de metodologia científica*. Atlas.
- [6] Muszinski, A.; Bertagnolli, S. de C. 2009. *Business Intelligence: um sistema de apoio a decisões gerenciais*.
- [7] Negash, S. 2004. *Business intelligence*. *Communications of the Association for Information Systems*, v. 13, p. 177 –195.
- [8] Schlegel, K. et al.. 2013. *Magic quadrant for business intelligence and analytics platforms*
- [9] Shi, C.; Zhang, D. 2010. Web service-based business intelligence system research and implementation. In: *Innovative Computing Information and Control*, 2008. p. 200–204.
- [10] Turban, E. et al. 2010. *Tecnologia da Informação para Gestão: Transformando os Negócios na Economia Digital*. Bookman.
- [11] Xu, L. et al. 2007. Research on business intelligence in enterprise computing environment. In: *Systems, Man and Cybernetics*, 2007. ISIC. IEEE International Conference.
- [12] English, B., Alderman, B., and Ferraz, M. 2010. *Microsoft SharePoint 2010 Administrator's Companion*. Microsoft Press, 1st Edition.
- [13] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. *From data mining to knowledge discovery: an overview*. In *Advances in knowledge discovery and data mining*, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthrusamy (Eds.). American Association for Artificial Intelligence, Menlo Park, CA, USA 1-34.
- [14] Avraham Leff and James T. Rayfield. 2001. *Web-Application Development Using the Model/View/Controller Design Pattern*. In *Proceedings of the 5th IEEE International Conference on Enterprise Distributed Object Computing (EDOC '01)*. IEEE Computer Society, Washington, DC, USA, 118-.
- [15] BRASIL,. 2004. Lei de nº 10.861, artigo 11, 14 de Abril de 2004. Disponível em http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/110.861.htm acesso em 04/03/2015

Mineração de regras de associação temporais quantitativas por meio de algoritmo genético

Alternative title: Quantitative temporal association rule mining by genetic algorithm

Sérgio F. da Silva^{*}, Marcos A. Batista[†]
Instituto de Biotecnologia – IBIOTEC/UFPA
Av. Dr. Lamartine P. Avelar, 1120
CEP: 75704-020, Catalão, GO
sergio@ufg.br, marcos.batista@cnpq.br

Agma J. M. Traina[‡]
Instituto de Ciências Matemáticas e de
Computação – ICMC/USP
Av. Trabalhador São-carlense, 400
CEP: 13566-590, São Carlos, SP
agma@icmc.usp.br

RESUMO

Mineração de regras de associação tem mostrado grande potencial para extrair conhecimento de conjunto de dados multidimensionais. Contudo, os métodos existentes na literatura não são efetivamente aplicáveis a dados temporais quantitativos. Este artigo estende os conceitos de mineração de regras de associação da literatura. Com base nestes conceitos é apresentado um método para mineração de regras de conjuntos de dados multidimensionais temporais quantitativos por meio de algoritmo genético, denominado GTARGA em referência à *Quantitative Temporal Association Rule Mining by Genetic Algorithm*. Experimentos com QTARGA em várias bases de dados reais mostram que este permite minerar várias regras de alta confiança em uma única execução do método.

Palavras-Chave

Algoritmos genéticos, mineração de regras de associação, dados temporais quantitativos.

ABSTRACT

Association rule mining has shown great potential to extract knowledge from multidimensional data sets. However, exist-

^{*}Sérgio F. da Silva é o autor principal da pesquisa.

[†]Marcos A. Batista contribuiu no desenvolvimento dos algoritmos desenvolvidos.

[‡]Agma J. M. Traina é a supervisora da pesquisa.

^{*}Sérgio F. da Silva é o autor principal da pesquisa.

[†]Marcos A. Batista contribuiu no desenvolvimento dos algoritmos desenvolvidos.

[‡]Agma J. M. Traina é a supervisora da pesquisa.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

ing methods in the literature are not effectively applicable to quantitative temporal data. This article extends the concepts of association rule mining from the literature. Based on the extended concepts is presented a method to mine rules from multidimensional temporal quantitative data sets using genetic algorithm, called GTARGA, in reference to *Quantitative Temporal Association Rule Mining by Genetic Algorithm*. Experiments with QTARGA in four real data sets show that it allows to mine several high-confidence rules in a single execution of the method.

Categorias e Descritores do Assunto

Intelligent Information Systems [**Data Mining**]: Genetic algorithms, Association rules mining, Temporal quantitative data

Termos Gerais

Concepts and methods

Keywords

Genetic algorithms, association rules mining, temporal quantitative data.

1. INTRODUÇÃO

Muitos fenômenos do mundo real, incluindo atividades e processos, apresentam variáveis correlacionadas. Desta forma, fenômenos reais podem ser melhor compreendidos por descobrir implicações dos valores de variáveis ao longo do tempo, ou seja, implicações de certos episódios em episódios subsequentes. Por exemplo, quebra de safras não são completamente compreendidas sem a análise de variáveis meteorológicas. Chuva ácida não é compreendida sem a análise de poluentes na atmosfera. Também, as implicações entre episódios são úteis para a geração de previsões (predição). Por exemplo, se tem-se que determinados poluentes na atmosfera acarretam chuva ácida e quantidade de tais poluentes tem aumento significativamente no decorrer dos últimos anos, então pode-se prever maior incidência de chuva ácida em um futuro próximo, com alta confiança.

Fontes de dados temporais quantitativos oriundas de atividades do mundo real (ou áreas do conhecimento) são oni-

presentes. Dentre algumas destas fontes pode-se citar: economia, comunicações, astronomia, energia, agronomia, meteorologia and agrometeorologia. Contudo, os dados produzidos por estas fontes têm sido praticamente inúteis devido a ausência de técnicas para extrair conhecimentos concretos e não-triviais destes. Atualmente, muitos destes dados são analisados por métodos estatísticos e/ou gráficos que têm capacidade limitada para a análise de múltiplas variáveis simultaneamente.

Nos últimos anos, vários métodos de mineração de dados temporais têm sido propostos [1, 2, 3, 4, 5, 7, 10, 11, 15, 17, 18, 19], assim como de mineração de dados quantitativos [8, 6, 12, 13, 14, 20]. Contudo, os métodos existentes na literatura não são efetivamente aplicáveis para a mineração de regras (implicações) de dados temporais quantitativos, devido a deficiências em satisfazer os seguintes critérios:

- Minerar regras (implicações temporais entre variáveis) e não somente identificar padrões em uma série temporal.
- Considerar o tempo de ocorrência de episódios e não somente a ordem em que eles ocorrem;
- Operar sem a necessidade do usuário informar parâmetros críticos tais como, limiares de suporte e de confiança das regras pretendidas;
- Minerar várias regras com diversidade e de alta confiança em uma única execução do método;
- Apresentar escalabilidade computacional, de modo a possibilitar sua aplicação para grandes conjuntos de dados.

Métodos da literatura atual não têm a capacidade de mineração regras como:

$\langle \langle \text{Prec. Acumulada (mm)} = [150, 220], \text{mês} = [\text{Nov}, \text{Dez}] \rangle \rangle$
 AND
 $\langle \text{Temperatura Média (}^{\circ}\text{C)} = [25, 32], \text{mês} = [\text{Nov}, \text{Dez}] \rangle \rangle$
 $\Rightarrow$
 $\langle \text{Crescimento de Planta (cm)} = [40, 50], \text{mês} = [\text{Dez}, \text{Jan}] \rangle \rangle,$

a qual provê conhecimento útil sobre a produção da cana-de-açúcar. Tais tipos de regras de associação são mineradas pelo método proposto neste artigo.

O método proposto, de agora em diante denominado QTARGA, em referência à *Quantitative Temporal Association Rule Mining by Genetic Algorithm*, se baseia em um algoritmo genético de código real. Algoritmos genéticos de código real têm sido bastante aplicados para mineração de regras de associação de dados quantitativos. Uma das principais dificuldades do uso de algoritmos genéticos para mineração de regras de associação, a qual tem sido ignorada em várias pesquisas, é a natureza unimodal destes algoritmos [16]. Algoritmos genéticos tradicionais são projetados para encontrar uma solução ótima ou sub-ótima. Desta forma, a população (conjunto de soluções candidatas) tende a concentrar em torno da solução ótima no decorrer das gerações (iterações) do algoritmo. Para solucionar este problema, foi usado um mecanismo de preservação de diversidade que permite ao AG explorar simultaneamente várias regiões do espaço de busca e, conseqüentemente, encontrar várias soluções ótimas

e/ou sub-ótimas em uma única execução do método. Vale destacar que o método de preservação de diversidade, a saber, *clearing* [16], é amplamente conhecido pela comunidade de algoritmos genéticos. Contudo, para viabilizar a utilização de *clearing* foi proposta uma medida de distância entre regras, sendo a distância entre duas regras proporcional à diversidade. Além disso, ao contrário de muitos métodos da literatura, QTARGA não usa de discretização prévia dos dados em intervalos, por ter mecanismos para descobrir tais intervalos no decorrer do processo de mineração.

O método QTARGA foi validado através de experimentos em quatro conjuntos de dados reais. Os resultados obtidos mostram que QTARGA consegue minerar várias regras com alta diversidade e confiança em uma única execução do método. Além disso, QTARGA apresenta boa escalabilidade computacional, o que permite sua aplicação para bases de dados grandes.

O restante deste artigo é organizado da seguinte forma. A Seção 2 define o formato das regras de interesse e a medida de qualidade destas. A Seção 3 apresenta o método QTARGA, desenvolvido para a mineração de regras temporais quantitativas, definidas conforme especificado na Seção 2. A Seção 4 reporta e discute os resultados obtidos pelo método QTARGA em quatro conjuntos de dados. Por fim, a Seção 5 apresenta as conclusões.

2. CONSIDERAÇÕES INICIAIS E DEFINIÇÕES

Este trabalho amplia a noção de regras de associação temporais apresentada em [9] para lidar efetivamente com dados temporais quantitativos. Os conceitos apresentados em [9], os quais formam uma das bases principais para mineração de regras de associação temporais, lidam somente com variáveis binárias (ou discretas), ex., *choveu/não_choveu* em um instante de tempo específico. Este tipo de representação é obviamente não representativo o suficiente para uma análise temporal quantitativa, pois, a intensidade de cada evento é desconsiderada. Buscando contornar essa limitação, foi proposto o conceito de *episodeset* como uma extensão do conceito de *itemset* para o domínio de dados temporais quantitativos. A seguir é definido o conceito de *episodeset*, além de conceitos adicionais que possibilitam a mineração de regras de associação temporais quantitativas.

DEFINIÇÃO 1. Episodeset. Seja $V = \{v_1, v_2, \dots, v_m\}$ o conjunto de variáveis de observação. O intervalo entre as observações é denominado granularidade de tempo (τ), a qual pode ser diária, semanal, mensal, entre outras. Seja $E = \{e_1, e_2, \dots, e_m\}$ um conjunto de episódios registrados no instante de tempo t , onde e_i é um episódio associado à variável v_i . Um *episodeset* $E^{(t)}$, $1 \leq t \leq n$ (onde n é o número de períodos de tempo do banco de dados temporal), registrado na base de dados $\mathcal{D}$ é chamado de *super-episodeset* ou *m-episodeset* pois, supõe-se que todas as variáveis de observação têm seus valores registrados¹. Assim, considera-se

¹Na coleta de dados do mundo real, por falhas de equipamentos ou humanas, pode acontecer de variáveis não terem determinados valores registrados, resultando em valores ausentes, ou porventura, os valores podem ser sido registrado de forma distorcida. Neste caso, normalmente utiliza-se técnicas de tratamento de valores ausentes (ou distorcidos), descarta-se aquele período ou, até mesmo, a variável.

que uma base de dados temporal quantitativa $\mathcal{D}$ é um conjunto de n super-episódios E , correspondentes a n períodos de tempo.

DEFINIÇÃO 2. Ciclo base. Um ciclo base Φ_i é uma sequência contígua de *super-episodesets* da base de dados. O comprimento (ou duração) l de um ciclo base é dado pelo número de *super-episodesets* por ciclo base, sendo este pré-estabelecido. Matematicamente, cada ciclo base corresponde ao intervalo de tempo $[t_{i,l}, t_{(i+1),l})$. O número de ciclos base n_{CB} é dado por $\lfloor |\mathcal{D}|/l \rfloor$, onde $|\mathcal{D}|$ é o número de *super-episodesets* da base de dados.

DEFINIÇÃO 3. Regra de associação temporal quantitativa. Uma regra de associação temporal quantitativa é uma implicação da forma $\mathbf{X} \Rightarrow \mathbf{Y}$ (se $\mathbf{X}$ então $\mathbf{Y}$), onde $\mathbf{X}$ e $\mathbf{Y}$ são conjunções de episódios associados às variáveis de observação. $\mathbf{V}_\mathbf{X}$ denota o subconjunto de variáveis associadas às condições episódicas de $\mathbf{X}$ e $\mathbf{V}_\mathbf{Y}$ denota o subconjunto de variáveis associadas com as condições episódicas $\mathbf{Y}$. A interseção de $\mathbf{V}_\mathbf{X}$ e $\mathbf{V}_\mathbf{Y}$ deve ser vazia, ie., $\mathbf{V}_\mathbf{X} \cap \mathbf{V}_\mathbf{Y} = \emptyset$.

DEFINIÇÃO 4. Condição episódica. Uma condição episódica é uma condição intervalar de uma variável em um dado intervalo de tempo. Condições episódicas são representadas da forma : $\langle v_i(\text{unidade de } v_i) = [v_{0i}, v_{1i}]$ no período (unidade de tempo) = $[t_0, t_1]$.

DEFINIÇÃO 5. Suporte de uma conjunção de condições episódicas. Uma conjunção de condições episódicas é dada por condições episódicas conectadas por pelo operador lógico AND. O suporte de uma conjunção de condições episódicas $\mathbf{X}$ representa a frequência de ocorrência de $\mathbf{X}$ na base de dados. Considerando que $\mathbf{X}$ é analisado por ciclo base, o suporte de $\mathbf{X}$ corresponde a frequência em que $\mathbf{X}$ nos ciclos base. Os *super-episodesets* de um ciclo base Φ_i , denotado por $\mathbf{E}[i]$, suportam uma conjunção de condições episódicas $\mathbf{X}$ se e somente se todas as condições episódicas de $\mathbf{X}$ ocorre em $\mathbf{E}[i]$. Matematicamente, o suporte é de $\mathbf{X}$ é definido como:

$$\text{supp}(\mathbf{X}) = \frac{\sum_{i=1}^{n_{BC}} \text{happen}(\mathbf{X}, \mathbf{E}[i])}{n_{BC}}, \quad (1)$$

onde $\text{happen}(\mathbf{X}, \mathbf{E}[i])$ retorna 1 se $\mathbf{X}$ ocorre em $\mathbf{E}[i]$ e 0, caso contrário.

3. O MÉTODO QTARGA

O método de mineração de regras de associação temporais quantitativas proposto se baseia na proposição de um algoritmo genético específico para o problema. Assim, o método é descrito em termos dos passos e operadores do algoritmo genético, o qual é dado no Algoritmo 1.

a) Codificação de cromossomo

Na codificação de cromossomo, cada variável da base de dados é associada a um gene. Considerando m variáveis de observação, tem-se m genes: $G_1, G_2, G_3, \dots, G_m$. Cada gene G_i de um cromossomo representa um episódio relacionado a variável $v_i, i = 1 \dots m$ e é codificado conforme ilustrado na Figura 1. Na Figura 1, w é um peso que é comparado a um limiar para indicar se a condição episódica representada pelo gene fará ou não parte da regra. AC é um *flag* para indicar se a condição episódica representada pelo

Algoritmo 1: Método QTARGA.

Entrada: Codificação de cromossomo, função de aptidão, restrições (atributos antecedentes / consequentes), comprimento de intervalo de atributo máximo ($piv.(\max v_i - \min v_i)$), janela de tempo mínima ($janTime$)

Saída: Regras de associação temporais quantitativas correspondentes a ótimos locais e globais.

- 1: Gere uma população de cromossomos ($\mathcal{C}$) aleatoriamente, de acordo com a codificação de cromossomo;
- 2: Avalie cada cromossomo $\mathcal{C}$ da população, conforme a função de aptidão;
- 3: Aplique o método de *niching*;
- 4: Selecione os cromossomos pelo método da roleta até completar o conjunto de pais (*matting pool*);
- 5: Aplique *crossover* uniforme tomando pares de indivíduos do *matting pool*;
- 6: Aplique mutação uniforme aos cromossomos recém gerados;
- 7: Selecione os melhores cromossomos entre e pais e filhos para a próxima geração;
- 8: Enquanto o número máximo de gerações não for atingido, retorne ao passo 2.
- 9: Retorne o conjunto de regras de associação codificadas pela população de cromossomos.

gene fará parte do antecedente ($flag = 0$) ou do consequente ($flag = 1$) da regra. v_0 e v_1 são os limites inferior e superior do intervalo da variável, respectivamente. t_0 e t_1 são os limites inferior e superior do intervalo de tempo, respectivamente. Os valores v_0, v_1, t_0 e t_1 são ajustados pelo algoritmo genético, respeitando as restrições das pelos parâmetros *piv* e *janTime*, que definem os tamanhos máximos de intervalos permitidos nas regras, para valores de variáveis e de tempo, respectivamente.

w	AC	v_0	v_1	t_0	t_1
-----	------	-------	-------	-------	-------

Figura 1: Representação de gene.

b) Medida de aptidão

$$\text{RelativeConfidence}(\mathbf{X} \Rightarrow \mathbf{Y}) =$$

$$\frac{\text{supp}(\mathbf{X} \cup \mathbf{Y}) - \text{supp}(\mathbf{X}) \text{supp}(\mathbf{Y})}{\text{supp}(\mathbf{X})(1 - \text{supp}(\mathbf{Y}))} \quad (2)$$

sendo $\text{supp}(\mathbf{Z})$ calculado conforme a Equação 1.

A *Relative Confidence* (confiança relativa) mede o grau de relação entre $\mathbf{X}$ e $\mathbf{Y}$. Ela retorna valores máximos quando $\mathbf{X}$ e $\mathbf{Y}$ ocorrem simultaneamente. Esta medida de qualidade de regras foi proposta recentemente na literatura [18] e vem sendo aplicada com sucesso para a mineração de regras interessantes. Contudo, o Método QTARGA funciona para qualquer medida quantitativa de qualidade de regras de associação.

c) Operadores genéticos

A seleção para reprodução é feita por meio do método da roleta. Pares de indivíduos selecionados para reprodução são cruzados por meio de *crossover* uniforme: sorteia-se uma máscara do tamanho do cromossomo, que indica qual cromossomo pai fornecerá cada gene ao primeiro filho; o segundo filho é gerado pelo complemento da máscara. Cada

cromossomo selecionado para mutação terá um de seus genes mudados. A mutação pode ocorrer no peso w , em AC (quando não for aplicadas restrições sobre as variáveis que compõem o antecedente e o consequente), ou nos limites inferiores (v_0 e t_0) e superiores (v_1 e t_1) dos intervalos de variáveis e de tempo.

d) *Niching*

Buscando realizar uma otimização multimodal, onde busca-se por vários ótimos locais e globais simultaneamente, foi usado um método de *niching*, denominado *clearing*, descrito em [16]. Para usar esse método, foi proposta uma medida de distância em regras de associação temporais quantitativas. A seguir é descrito o método de *clearing niching* e a medida de distância proposta.

O método de *clearing* corresponde ao conceito de *niching* enunciado por J. H. Holland em 1975: o compartilhamento de recursos por uma população de indivíduos caracterizados por alguma similaridade. Porém, ao invés de compartilhar os recursos disponíveis, o método de *clearing* provê os recursos de um nicho somente ao melhor indivíduo de cada subpopulação. Isto permite ao algoritmo genético realizar uma otimização multimodal. Além disso, o método de *clearing* permite ao AG reduzir o problema de deriva genética (*genetic drift*) [16], quando usado em conjunto com um operador de seleção apropriado.

Clearing é aplicado entre a avaliação da aptidão dos cromossomos e a aplicação do operador de seleção para cruzamento. O método faz uso de uma medida de distância (dissimilaridade) entre cromossomos (cujo fenótipo corresponde a regras de associação) para determinar se eles/(elas) pertencem a uma mesma subpopulação ou não. Cada subpopulação terá um cromossomo dominante: o que tem o maior valor de aptidão na subpopulação. Se um cromossomo pertence a uma subpopulação, então sua dissimilaridade com relação ao cromossomo dominante é menor que um dado limiar σ , denominado raio de *clearing*. O método de *clearing* preserva a aptidão do cromossomo dominante enquanto que diminui para zero a aptidão dos demais cromossomos da população. Assim, o método de *clearing* atribui todos os recursos de um nicho para um único cromossomo: o vencedor (*winner*). Tal método corresponde a remover imaginariamente da população todos os indivíduos dominados dentro de seus nichos.

Também, o método de *clearing* é generalizável para aceitar vários vencedores, escolhidos entre os melhores indivíduos do nicho [16]. A capacidade de um nicho é definida como o número máximo de cromossomos que um nicho pode comportar. Se a capacidade de nicho for igual ao tamanho da população o efeito de *clearing* desaparece e o método de busca torna-se um GA padrão. A escolha de capacidade de nicho entre 1 e o tamanho da população oferece situações intermediárias entre o efeito de *clearing* máximo e uma busca GA padrão.

O algoritmo 2 descreve o método de *clearing* [16]. Considere $\mathcal{C}$ (população de cromossomos) e n_c (número de cromossomos da população) como sendo variáveis globais. σ é o raio de *clearing* e κ é a capacidade de cada nicho. A variável *nbWinner* armazena o número de vencedores do nicho corrente. A população de cromossomos $\mathcal{C}$ é considerada como sendo um vetor de n_c cromossomos.

O algoritmo de *clearing* (Algoritmo 2) usa três funções:

- *OrdenaFitness*($\mathcal{C}$): ordena a população de cromosso-

Algoritmo 2: *Clearing niching*.

Entrada: σ (raio de *clearing*), κ (capacidade de cada nicho).
Saída: *Clearing* – atribuição dos recursos de um nicho ao indivíduo mais apto.

```

1: OrdenaFitness( $\mathcal{C}$ );
2: para  $i = 0$  até  $n_c - 1$ 
3:   se Fitness( $\mathcal{C}[i]$ ) > 0
4:     nbWinners = 1;
5:     para  $j = i + 1$  até  $n_c - 1$ 
6:       se Fitness( $\mathcal{C}[j]$ ) > 0 AND Distância(f( $\mathcal{C}[i]$ ),
7:         f( $\mathcal{C}[j]$ )) <  $\sigma$ 
8:         se nbWinners <  $\kappa$ 
9:           nbWinners = nbWinners + 1;
10:        senão Fitness( $\mathcal{C}[j]$ ) = 0;
10:        Fitness( $\mathcal{C}[j]$ ) = 0;
```

mos em ordem decrescente de acordo com a aptidão.

- *Fitness*($\mathcal{C}[i]$): retorna a aptidão do i -ésimo cromossomo da população $\mathcal{C}$.
- *Distância*(*f*($\mathcal{C}[i]$), *f*($\mathcal{C}[j]$)): retorna a distância entre o fenótipo de dois cromossomos da população;
- *f*($\mathcal{C}[i]$): retorna o fenótipo do cromossomo $\mathcal{C}[i]$, o qual é uma regra de associação temporal quantitativa no presente trabalho.

e) Calculando a distância entre regras de associação

Sejam duas regras de associação:

$$\begin{aligned}
 R &= (CA_R(v_{j_1}) \text{ AND } \dots \text{ AND } CA_R(v_{j_n})) \Rightarrow \\
 & (CC_R^d(v_{j_1}) \text{ AND } \dots \text{ AND } CC_R(v_{j_m})) \text{ e} \\
 S &= (CA_S(v_{j_1}) \text{ AND } \dots \text{ AND } CA_S(v_{j_o})) \Rightarrow \\
 & (CC_S^d(v_{j_1}) \text{ AND } \dots \text{ AND } CC_S(v_{j_p})), \text{ onde } CA_R(v_{j_i}), \\
 & CC_R(v_{j_i}), CA_S(v_{j_i}), CC_S(v_{j_i}) \text{ são condições episódicas} \\
 & \text{relacionadas às variáveis } v_{j_i} \text{ da base de dados. Cada} \\
 & \text{condição episódica tem a forma:}
 \end{aligned}$$

$$v_i \in [v_{0i}, v_{1i}] \text{ no intervalo de tempo } [t_{0i}, t_{1i}]$$

onde v_i é uma variável qualquer da base de dados. Para uma dada regra R , a interseção das variáveis associadas às condições episódicas do antecedente (VA_R) com as variáveis associadas às condições episódicas do consequente (VC_R) deve ser vazia, isto é, $VA_R \cap VC_R = \emptyset$.

A distância entre R e S , denotada por *Distance*(R, S) é dada pelo algoritmo 3. Para cada condição episódica no antecedente da regra R , CA_R , é verificado se existe alguma condição episódica no antecedente da regra S , CA_S , que seja comparável com CA_R . Duas condições episódicas são comparáveis se elas se referem à mesma variável. Se existe duas condições episódicas comparáveis, calcula-se a distância entre elas. Senão incrementa-se 1 (um) no contador de distância. Em seguida o contador de distância é dividido pelo número de condições episódicas n_{CA_R} . O mesmo cálculo é feito para as condições episódicas do consequente. Finalmente a distância é entre as regras é dada por $(dist_A + dist_C)/2$.

A distância entre condições episódicas, *DistanciaEp*($C1, C2$), é dada pelo Algoritmo 4.

4. RESULTADOS

Nesta seção são reportados quatro casos de estudo realizados para a validação da técnica desenvolvida. Todos os

Algoritmo 3: $Distância(R, S)$ – Distância entre duas regras de associação temporais quantitativas.

Entrada: R, S (duas regras de associação temporais quantitativas).
Saída: distância entre R e S (dist).
1: $n_{CA_R} = numCond(CA_R)$;
2: $dist_A = 0$;
3: **para** $i = 0$ **até** $n_{CA_R} - 1$
4: **se** $\exists CA_S(v_{j_k})$ tal que $v_{j_k} == v_{j_i}, v_{j_i} \in VA_R$
5: $dist_A = dist_A + DistanciaEp(CA_R(v_{j_i}), CA_S(v_{j_k}))$;
6: **senão**
7: $dist_A = dist_A + 1$;
8: $dist_A = dist_A / n_{CA_R}$;

9: $n_{CC_R} = numCond(CC_R)$;
10: $dist_C = 0$;
11: **para** $i = 0$ **até** $n_{CC_R} - 1$
12: **se** $\exists CC_S(v_{j_k})$ tal que $v_{j_k} == v_{j_i}, v_{j_i} \in VC_R$
13: $dist_C = dist_C + DistanciaEp(CC_R(v_{j_i}), CC_S(v_{j_k}))$;
14: **senão**
15: $dist_C = dist_C + 1$;
16: $dist_C = dist_C / n_{CC_R}$;
17: $dist = (dist_A + dist_C) / 2$;

experimentos foram processados em um MacBook Pro, Processador 2.9 GHz Intel Core i7 com 8GB de memória DDR3 1600 MHz usando o sistema operacional OS X 10.8.3. Os métodos foram implementados em linguagem C. Nos casos de estudo foram analisados o número de regras mineradas ($numRules$) e o tempo de execução ($exec. time$) dado em segundos, em função de configurações da técnica. Na configuração da técnica foi considerado variações no tamanho de população de cromossomos (n_C), o número de gerações ($nGen$), o raio de *clearing* (σ) e o tamanho máximo de intervalo de variável admissível é definido por $piv \cdot (\max v - \min v)$, onde piv é uma porcentagem. Nos resultados é reportado somente o valor de piv utilizado. Os demais parâmetros do AG, mantidos fixos nos casos de estudo, são: taxa de cruzamento de 80% e taxa de mutação de 3% por cromossomo. Os resultados reportados corresponde a uma média de três execuções para cada configuração da técnica.

4.1 Caso de estudo 1: Stock Prices

Neste caso de estudo foi utilizada a base de dados *Stock Prices* (<http://www.stat.ucla.edu/cases/>), que corresponde aos preços de ações diários de janeiro de 1988 à outubro de 1991, de dez companhias aéreas dos Estados Unidos (dez variáveis, pois o preço de ações de cada companhia é uma variável). Neste caso, o ciclo base é semanal, ou seja, $l = 7$.

Neste caso de estudo foi imposta a restrição que a companhia10 (*company10*) fará parte do consequente da regra. A tabela 1 mostra os resultados obtidos em termos da configuração da técnica (especificada no início da seção de resultados), do número de regras mineradas e do tempo de execução. As conclusões gerais dos experimentos são dadas na subseção 4.5.

A Figura 2 mostra algumas regras mineradas para a base de dados *Stock Prices*. Pode-se verificar uma correlação entre os preços de ações das companhias 3 e 10: quando o preço de ações da companhia 3 (*company3*) é relativamente alto, o preço de ações da companhia 10 (*company10*) é rela-

Algoritmo 4: $DistanciaEp(C_1, C_2)$ – Distância entre dois episódios temporais associados à mesma variável.

Entrada: C_1, C_2, v (duas condições episódicas associadas a uma mesma variável v).
Saída: distância entre C_1 e C_2 , representada por ($distEp$).
1: calcular:
 $\min v$ /*valor mínimo que v assume na base de dados*/
 $\max v$ /*valor máximo que v assume na base de dados*/
 $\min t$ /*valor mínimo que t assume na base de dados*/
 $\max t$ /*valor máximo que t assume na base de dados*/
2: $dv = \min\{(v_1^{(C_1)} - v_0^{(C_1)}), (v_1^{(C_2)} - v_0^{(C_2)})\} - (\min\{v_1^{(C_1)}, v_1^{(C_2)}\} - \max\{v_0^{(C_1)}, v_0^{(C_2)}\})$;
3: $distV = dv / (\max v - \min v)$;

4: $dt = \min\{(t_1^{(C_1)} - t_0^{(C_1)}), (t_1^{(C_2)} - t_0^{(C_2)})\} - (\min\{t_1^{(C_1)}, t_1^{(C_2)}\} - \max\{t_0^{(C_1)}, t_0^{(C_2)}\})$;
5: $distT = dt / (\max t - \min t)$;
6: $distEp = (distV + distT) / 2$;

tivamente baixo; e quando o preço de ações da companhia 3 (*company3*) é relativamente baixo, o preço de ações da companhia 10 (*company10*) é relativamente alto.

n_C	$nGen$	σ	piv	$numRules$	$exec. time (sec.)$
100	350	0.3	0.1	44	4.18
100	350	0.3	0.2	36	3.76
100	350	0.5	0.1	63	4.05
100	350	0.5	0.2	54	4.20
50	250	0.3	0.1	28	0.96
50	250	0.3	0.2	23	0.90
50	250	0.5	0.1	23	0.83
50	250	0.5	0.2	31	0.98

Tabela 1: Configuração do AG *versus* número de regras mineradas ($numRules$) *versus* tempo de execução ($exec. time$) em segundos para a base de dados *Stock Prices*.

4.2 Caso de estudo 2: Araraquara

Neste caso de estudo foi utilizada a base de dados denominada Araraquara, coletada pelo Sistema de Monitoramento Agrometeorológico – Agritempo (<http://www.agritempo.gov.br/>). Ela contém dados agrometeorológicos mensais de Araraquara correspondentes aos valores de média da temperatura mínima ($Tmin$), média da temperatura máxima ($Tmax$), precipitação acumulada ($Prec$), *Normalized Difference Vegetation Index* (NDVI) médio e *Water Requirement Satisfaction Index* (WRSI) médio. Os dados correspondem ao período de abril de 2001 a janeiro de 2008. Neste caso, o ciclo base é anual, ou seja, $l = 12$.

Neste caso de estudo foi imposta a restrição que o NDVI fará parte do consequente da regra. A tabela 2 sumariza os resultados obtidos.

4.3 Caso de estudo 3: Piracicaba

Neste caso de estudo foi utilizada a base de dados Piracicaba, coletada pela Embrapa (Empresa Brasileira de Pesquisa Agropecuária). A base de dados consiste de três va-

Rule:	company1 (US\$) in [55.68, 61.50] on Tuesdays-Wednesday ==> company10 (US\$) in [54.65, 60.25] on Tuesdays-Wednesday Fitness = 1.000
Rule:	company3 (US\$) in [22.68, 23.92] on Thursdays-Fridays ==> company10 (US\$) in [39.40, 42.20] on Thursdays-Fridays Fitness = 1.000
Rule:	company3 (US\$) in [16.80, 18.04] on Mondays-Tuesdays ==> company10 (US\$) in [43.96, 46.76] on Mondays-Tuesdays Fitness = 1.000
Rule:	Rule: company5 (Stock_Price) in [70.16, 76.80] on Thursdays-Fridays AND company6 (Stock_Price) in [22.66, 24.77] on Wednesday-Thursdays ==> company10 (Stock_Price) in [53.95, 56.75] on Thursdays-Fridays Fitness = 1.000 Fitness = 1.000

Figura 2: Exemplo de regras mineradas para a base de dados *Stock Prices*.

nc	$nGen$	σ	piv	$numRules$	$exec. time (sec.)$
100	350	0.3	0.1	11	0.49
100	350	0.3	0.2	17	0.63
100	350	0.5	0.1	4	0.26
100	350	0.5	0.2	6	0.29
50	250	0.3	0.1	8	0.17
50	250	0.3	0.2	10	0.16
50	250	0.5	0.1	4	0.11
50	250	0.5	0.2	4	0.12

Tabela 2: Configuração AG *versus* número de regras mineradas ($numRules$) *versus* tempo de execução ($exec. time$) em segundos para a base de dados Piracicaba.

riáveis tomadas mensalmente: o valor médio de temperatura máxima ($Tmax$), o valor médio de temperatura mínima ($Tmin$), e a precipitação acumulada ($Prec$). Os dados correspondem a um período de 47 anos (de 1961 a 2008). Neste caso, o ciclo base é anual, ou seja, $l = 12$.

Neste caso de estudo foi imposta a restrição que a precipitação ($Prec$) fará parte do consequente da regra. A tabela 3 sumariza os resultados obtidos.

4.4 Caso de estudo 4: Produtividade da cana-de-açúcar em Piracicaba

Neste caso de estudo foi utilizada uma base de dados de produtividade da cana-de-açúcar em Piracicaba, fornecida pelo Cepagri (Centro de Pesquisas Agrometeorológicas e Climáticas Aplicadas à Agricultura)-UNICAMP. A base de dados consiste de quatro variáveis tomadas mensalmente no município de Piracicaba: média da temperatura mínima ($Tmin$), média da temperatura máxima ($Tmax$), precipitação média ($Prec$) e produtividade da cana-de-açúcar ($Prod$) em toneladas por hectare (ton/hect). A base de dados corresponde ao período de 2003 a 2009. Também neste caso, o

nc	$nGen$	σ	piv	$numRules$	$exec. time (sec.)$
100	350	0.3	0.1	11	0.49
100	350	0.3	0.2	17	0.63
100	350	0.5	0.1	4	0.26
100	350	0.5	0.2	6	0.29
50	250	0.3	0.1	8	0.17
50	250	0.3	0.2	10	0.16
50	250	0.5	0.1	4	0.11
50	250	0.5	0.2	4	0.12

Tabela 3: Configuração do AG *versus* número de regras mineradas ($numRules$) *versus* tempo de execução ($exec. time$) em segundos para a base de dados Piracicaba.

ciclo base é anual, ou seja, $l = 12$.

No presente caso de estudo foi imposta a restrição que a produtividade ($Prod$) fará parte do consequente da regra. A tabela 3 sumariza os resultados obtidos.

nc	$nGen$	σ	piv	$numRules$	$exec. time (sec.)$
100	350	0.3	0.1	10	0.29
100	350	0.3	0.2	14	0.34
100	350	0.5	0.1	4	0.12
100	350	0.5	0.2	8	0.20
50	250	0.3	0.1	9	0.12
50	250	0.3	0.2	8	0.12
50	250	0.5	0.1	4	0.06
50	250	0.5	0.2	5	0.08

Tabela 4: Configuração do AG *versus* número de regras mineradas ($numRules$) *versus* tempo de execução ($exec. time$) em segundos para a base de dados Produtividade da cana-de-açúcar em Piracicaba.

4.5 Análise dos resultados

No geral pode-se perceber que a técnica consegue minerar várias regras de associação, sendo elas diversas e de alta qualidade, de acordo com a medida de aptidão. Em média a aptidão das regras mineradas é próxima de 0.99, sendo que o valor máximo da função de aptidão (equação 2) é 1 (um). A configuração do tamanho de população (nc) e número de gerações ($nGen$) se comportou conforme o esperado: quando maior os valores destes parâmetros, maior o número de regras mineradas. Contudo, aumentando-se os valores de nc e $nGen$, aumenta-se o tempo de execução do método, conforme é esperado.

A intuição quanto raio de *clearing* é que, quanto maior este, maior é capacidade do ambiente para comportar nichos, e consequentemente, maior a quantidade de regras mineradas. A configuração deste parâmetro se comportou em média conforme o esperado, contudo o caso de estudo 1 foi uma exceção. Esta exceção aconteceu devido ao grande número de padrões (regras) existentes na base *Stoke Prices*. Este fato foi detectado aumentando o tamanho de população nc . Com populações a partir de 300 cromossomos, o ajuste do raio de *clearing* (σ) passa a comportar conforme o esperado: quanto menor o raio de *clearing*, maior o número de regras mineradas. O parâmetro piv também se comportou conforme o esperado; em média, quanto maior o tamanho de intervalo admissível, maior o número de regras mineradas. Contudo, se piv for muito grade (próximo de 1), serão mineradas regras que não expressam nenhum conhecimento concreto (específico). Por exemplo, temperatura ($^{\circ}C$) [0-42], sempre ocorre em Araraquara.

5. CONCLUSÕES

Partindo de conceitos existentes na literatura sobre mineração de regras de associações temporais, foi proposta uma extensão para lidar com dados temporais quantitativos, de forma natural e sem perda de informação. Para demonstrar a validade dos conceitos propostos, foi elaborado um algoritmo genético que funciona de acordo com tais conceitos de forma a identificar implicações temporais entre episódios. Vale destacar que o algoritmo genético proposto não necessita de discretização prévia dos dados em intervalos, assim como, da especificação de parâmetros críticos específicos de bases de dados como os limiares de suporte e confiança, necessários em algoritmos de mineração de regras de associação clássicos. Como pode ser notado, o algoritmo genético proposto tem vários parâmetros ajustáveis, contudo, o ajuste destes parâmetros é intuitivo, conforme descrito na subseção 4.5. A limitação clássica de algoritmos genéticos tradicionais, de realizar busca unimodal, foi tratada através da utilização do mecanismo clássico de *clearing* para preservação de diversidade. Os resultados obtidos mostram que é possível minerar várias regras de alta qualidade em uma única execução do método.

Também vale destacar que, conforme é de conhecimentos dos autores, o tipo de regras mineradas pela abordagem proposta, além de altamente informativa, não é identificada por nenhum método de mineração regras de associação da literatura. Assim, o método proposto não foi comparado com métodos da literatura. Espera-se que os conceitos propostos sejam uma base para definição de novos métodos de mineração de regras de associação temporais quantitativos. Além disso, o método proposto pode ser aplicado em bases de dados temporais quantitativos advindos de variadas áreas do conhecimento, de forma a descobrir conhecimentos valiosos e não-triviais.

6. AGRADECIMENTOS

Agradecemos a FAPEG (Proc. 201210267900) e ao CNPq (Proc. 447984/2014-4 e Proc. 479792/2012-7) pelo suporte financeiro.

7. REFERÊNCIAS

- [1] S. Amo, N. A. Silva, R. P. Silva, and F. S. Pereira. Tree pattern mining with tree automata constraints. *Information Systems*, 35:570–591, 2010.
- [2] B. Catania and A. Maddalena. A unified framework for heterogeneous patterns. *Information Systems*, 37:460–483, 2012.
- [3] D.-A. Chiang, C.-T. Wang, S.-P. Chen, and C.-C. Chen. The cyclic model analysis on sequential patterns. *IEEE Transactions on Knowledge and Data Engineering*, 21(11):1617–1628, 2009.
- [4] J. K. Febrer-Hernández and J. Hernández-Palancar. Sequential pattern mining algorithms review. *Intelligent Data Analysis*, 16:451–466, 2012.
- [5] T. Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.
- [6] E. Georgii, L. Richter, U. Rückert, and S. Kramer. Analyzing microarray data using quantitative association rules. *Bioinformatics*, 21(suppl 2):ii123–ii129, 2005.
- [7] Y. Hai and X. Li. A general temporal association rule frequent itemsets mining algorithm. *International Journal of Advancements in Computing Technology*, 3(11):63–71, 2011.
- [8] T.-P. Hong, C.-S. Kuo, and S.-C. Chi. Mining association rules from quantitative data. *Intelligent Data Analysis*, 3(5):363–376, 1999.
- [9] C.-H. Lee, M.-S. Chen, and C.-R. Lin. Progressive partition miner: an efficient algorithm for mining general temporal association rules. *IEEE Transaction on Knowledge and Data Engineering*, 15(4):1004–1017, 2003.
- [10] Y. J. Lee, J. W. Lee, D. J. Chai, B. H. Hwang, and K. H. Ryu. Mining temporal interval relational rules from temporal data. *Journal of Systems and Software*, 82(1):155–167, 2009.
- [11] N. R. Mabroukeh and C. I. Ezeife. A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys*, 43(1):1–41, 2010.
- [12] D. Martin, A. Rosete, J. Alcalá-Fdez, and F. Herrera. A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules. *Evolutionary Computation, IEEE Transactions on*, 18(1):54–69, 2014.
- [13] M. Martínez-Ballesteros, A. Troncoso, F. Martínez-Álvarez, and J. C. Riquelme. Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution. *Integrated Computer-Aided Engineering*, 17(3):227–242, 2010.
- [14] B. Pei, S. Zhao, H. Chen, X. Zhou, and D. Chen. Farp: Mining fuzzy association rules from a probabilistic quantitative database. *Information Sciences*, 237:242 – 260, 2013.
- [15] M. Plantevit, A. Laurent, D. Laurent, M. Teisseire, and Y. W. Choong. Mining multidimensional and multilevel sequential patterns. *ACM Transactions on Knowledge Discovery from Data*, 4(1):1–37, 2010.
- [16] A. Pétrowski. A clearing procedure as a niching method for genetic algorithms. In *Proceedings of 3rd IEEE International Conference on Evolutionary Computation*, pages 798–803, New York, USA, 1996.
- [17] N. Tatti and B. Cule. Mining closed strict episodes. *Data Mining and Knowledge Discovery*, 25(1):34–66, 2011.
- [18] X. Yan, C. Zhang, and S. Zhang. Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. *Expert Systems with Applications*, 36:3066–3076, 2009.
- [19] M. Zhang and C. He. Survey on association rules mining algorithms. *Advancing Computing, Communication, Control and Management*, 56:111–118, 2010.
- [20] H. Zheng, J. He, G. Huang, and Y. Zhang. Optimized fuzzy association rule mining for quantitative data. In *2014 IEEE International Conference on Fuzzy Systems*, pages 396–403, Beijing, China, 2014.

Aprendizado de Máquina em Sistemas de Recomendação Baseados em Conteúdo Textual: Uma Revisão Sistemática

Alternative Title: Machine Learning in Textual Content-Based Recommendation Systems: A Systematic Review

Lucas F. Brunialti
Universidade de São Paulo
São Paulo – SP – Brasil
lucas.brunialti@usp.br

Valdinei Freire
Universidade de São Paulo
São Paulo – SP – Brasil
valdinei.freire@usp.br

Sarajane M. Peres
Universidade de São Paulo
São Paulo – SP – Brasil
sarajane@usp.br

Clodoaldo A. M. Lima
Universidade de São Paulo
São Paulo – SP – Brasil
c.lima@usp.br

RESUMO

Sistemas de Recomendação baseados em Conteúdo (SRbC) é uma área em que estratégias de Aprendizado de Máquina (AM) podem ser potencialmente aplicadas com êxito. Contudo, especificamente na área de SRbC textual, o uso de AM não tem sido expressivo nos últimos anos. Neste artigo é apresentada uma Revisão Sistemática para identificação, interpretação e avaliação de como estratégias de AM vêm sendo utilizadas no contexto de SRbC textual a fim de contribuir para a evolução da interseção de tais áreas.

Palavras-Chave

Sistemas de Recomendação baseado em Conteúdo, Conteúdo Textual, Aprendizado de Máquina, Revisão Sistemática

ABSTRACT

Content-based Recommendation Systems (CbRS) is a research area in which Machine Learning (ML) strategies can be applied with success. However, specifically in textual CbRS, the use of ML has not been expressive in recent years. To contribute to the evolution of the intersection of such areas, we present a Systematic Review to identify, interpret and evaluate how the ML strategies have been applied to CbRS.

Categories and Subject Descriptors

H.4 [Information Systems]: Miscellaneous; I.2 [Artificial Intelligence]: Learning—*Induction, Knowledge Acquisition*

General Terms

Design, Theory

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

Keywords

Content-Based Recommendation Systems, Textual Content, Machine Learning, Systematic Review

1. INTRODUÇÃO

Um sistema de recomendação (SR) tem o objetivo de sugerir itens de forma a satisfazer sob algum aspecto as necessidades de um usuário. Geralmente, esses sistemas atuam em contextos onde a tomada de decisão sobre a escolha de itens se dá em um conjunto grande de opções, no qual uma busca por meio de mecanismos clássicos como palavras-chave ou termos de interesse tem a chance de retornar resultados insatisfatórios. Os primeiros SR surgiram no início da década de 1990 [1], e um dos primeiros sistemas propostos chamava-se *Tapestry* [20], o qual introduziu a *Filtragem Colaborativa* como uma técnica para implementação da recomendação.

Em [10] é proposta uma taxonomia para tipos básicos de SR, de acordo com a forma como a recomendação é implementada: *colaborativo*, no qual a recomendação se dá com base em itens com os quais usuários interagiram no passado, relacionando as avaliações de usuários sobre um mesmo item; *baseado em conteúdo* (SRbC), no qual as recomendações são geradas com base nas características dos itens e no perfil do usuário; *baseado em conhecimento*, em que sugestões de itens são baseadas em inferências sobre as necessidades e preferências dos usuários. Cada um destes tipos possui vantagens e desvantagens que podem ser superadas por meio de sistemas *híbridos*, os quais são implementados a partir da combinação das estratégias usadas nos tipos básicos.

Os SR são considerados ferramentas de grande utilidade, principalmente para usuários de sistemas *online* [21, 25], os quais se deparam com um montante de informação maior do que conseguem lidar adequadamente; a indústria de software para *web* já considera os SR como um meio de atender melhor as demandas dos clientes e tornar seus sistemas mais lucrativos [12]. Nesse contexto, o processamento de dados textuais (notícias, livros, artigos científicos) representa uma demanda importante e esse tipo de dado é, frequentemente, objeto de processamento de SRbC [5, 7, 13, 30].

Aprendizado de Máquina (AM) é caracterizado pelo de-

envolvimento de técnicas que objetivam prover os softwares com a habilidade de melhorar seu desempenho em uma tarefa aprendendo através da experiência (aprendizado indutivo) [15]. Diferentes técnicas de AM, ou combinações delas [3], têm sido usadas para análise de dados textuais em diversos contextos [6, 30]. O sucesso obtido nesses contextos mostra que tais técnicas são adequadas para o processamento de dados textuais e, portanto, são adequadas para a construção de SRs que lidam com este tipo de dado.

A presente Revisão Sistemática (RS) foi conduzida com o objetivo de construir um panorama sobre as pesquisas mais recentes na área de SRbC textual construídos com base em técnicas de AM. A RS é apresentada neste artigo como segue: Seção 2 apresenta os conceitos básicos de SRbC e AM; Seção 3 resume alguns trabalhos relacionados a esta RS; Seção 4 traz a metodologia de construção da RS; Seções 5 e 6 apresentam os resultados da revisão e uma discussão sobre eles; a conclusão é apresentada na Seção 7.

2. CONCEITOS FUNDAMENTAIS

Esta seção apresenta os conceitos básicos sobre SRbC e AM, para dar subsídios para o entendimento dos resultados e discussões apresentados das seções seguintes.

2.1 Sistemas de Recomendação baseados em Conteúdo

Para fazer uma recomendação, os SRbC analisam o conteúdo dos itens candidatos à recomendação por meio da extração de atributos que os descrevem e analisam o perfil do usuário, que pode ser ou não representado por meio dos itens que eles preferem. Conhecendo o perfil dos usuários e o conteúdo referente aos itens, o sistema deve ser capaz de recomendar adequadamente novos itens aos usuários.

Formalmente, seja um conjunto de usuários $\mathcal{U} = \{U_1, \dots, U_n, \dots, U_N\}$, um conjunto de itens $\mathcal{I} = \{I_1, \dots, I_m, \dots, I_M\}$ e o histórico de interações $\mathcal{H}$, no qual uma interação $h \in \mathcal{H}$ é uma tripla $h = (U_n, I_m, \sigma)$ indicando que o usuário U_n interagiu com o item I_m e emitiu a avaliação σ sobre a experiência. Define-se $\mathcal{H}_{U_n} \subset \mathcal{H}$ como o histórico de interações do usuário U_n . O problema de recomendação com base no conteúdo pode ser visto como a definição de uma relação de similaridade $s: \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R}$ que permita aproximar uma função $l_s: 2^{\mathcal{H}} \rightarrow 2^{\mathcal{I}}$, de tal forma que a lista de recomendações gerada para o usuário U_n é dada por $L_{U_n} = l_s(\mathcal{H}_{U_n})$.

Uma arquitetura típica para SRbC, apresentada em [14], é composta por três módulos (Figura 1). O módulo de análise e representação de conteúdo é responsável por receber as descrições dos itens, pré-processá-las a fim de analisar o conteúdo dos mesmos e criar uma representação que possibilite o aprendizado dos perfis dos usuários e a elaboração das recomendações. O módulo de personalização usa tal representação e o histórico de interações do usuário ($\mathcal{H}_{u_n}$) para induzir a função $l_s(\mathcal{H}_{u_n})$. As avaliações da interação entre usuário e item pode ser explícita, em que ele avalia um item, ou implícita, na qual a execução da interação e suas características representam uma avaliação (o usuário leu uma notícia recomendada, ou permaneceu por muito tempo em uma página). O terceiro módulo é composto pelo componente de filtragem. Basicamente esse componente utiliza a saída da função $l_s(\mathcal{H}_{u_n})$ para gerar uma lista de recomendações (L_{u_n}), e a apresenta ao usuário. Tipicamente a lista de recomendações é ordenada com base em um *score* e os itens mais relevantes (*top N*) são apresentados.

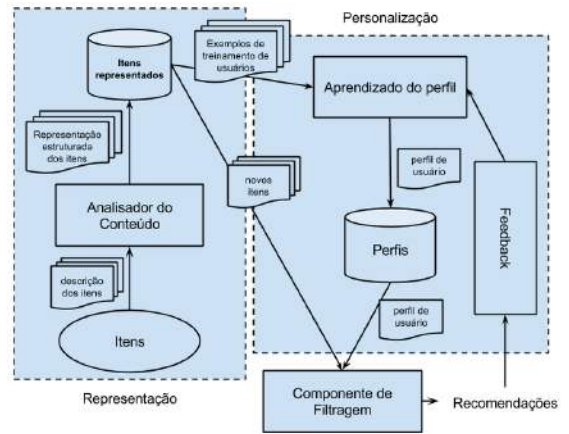


Figura 1: Arquitetura de um SRbC - adap. de [14]

Cada um dos tipos de SR tem vantagens e desvantagens. No caso de SRbC, as seguintes vantagens se destacam [1, 14]: o sistema não necessita de uma base de usuários, visto que suas recomendações são baseadas nas relações entre os itens; itens que nunca foram recomendados e avaliados por usuários (conhecido como problema de *cold-start* de item) podem aparecer na lista de recomendações, dado que é possível descobrir as suas relações com outros itens; e a recomendação de um item pode ser justificada, já que ela é baseada no conteúdo do item e na sua similaridade com itens já acessados pelo usuário. Já em termos de desvantagens, [1, 10, 14] destacam: a análise de conteúdo (textual ou multimídia) é uma tarefa complexa e exige o uso de técnicas sofisticadas para associar conteúdo a interesses de usuários no aprendizado dos perfis; dificuldade de recomendar itens que não pertençam a um mesmo domínio ou domínios similares (dentro do que o sistema aprendeu como perfil de usuário), mas que podem ser de interesse do usuário (conhecido como problema de serendipidade); e interesses de longo prazo para novos usuários podem não ser captados pelo sistema (conhecido como problema de *cold-start* de usuário), uma vez que para novos usuários há pouca informação sobre avaliações de itens.

Nesta RS são analisados estudos que implementam SRbC textual. O desenvolvimento desses SR possui um dificultador que é a necessidade de tratamento de um tipo de dado não estruturado. Isso implica na necessidade de uso de técnicas de pré-processamento de textos, técnicas de representação de texto estruturada, técnicas de redução de dimensionalidade e proposição de tópicos, além do que já é necessário para resolver o problema de recomendação. Detalhes sobre esses tópicos são encontrados em [22, 24, 26, 30].

2.2 Aprendizado de Máquina

A teoria de AM é baseada nos princípios do aprendizado indutivo (AI), ou seja, modelos são determinados a partir de um conjunto de dados ou representações de experiências [18]. Normalmente, o aprendizado indutivo é implementado por algoritmos que processam um conjunto de dados e extraem um modelo capaz de explicar ou representar os dados sob algum aspecto. Esse modelo pode ser usado para explicar ou representar um novo dado (do mesmo domínio do conjunto de dados inicial), que é apresentado *a posteriori*.

O AI pode ser de três modalidades: supervisionado, não-

supervisionado e semi-supervisionado. Na primeira, os algoritmos ajustam parâmetros de um modelo a partir do erro medido entre respostas obtidas e esperadas. Na segunda, os parâmetros de um modelo são ajustados com base na maximização de medidas de qualidade das respostas obtidas. A terceira é caracterizada pelo uso de algoritmos híbridos, que fazem uso dos recursos de correção de erro e de maximização de medidas de qualidade, conforme necessário.

3. TRABALHOS RELACIONADOS

Os trabalhos relacionados a esta RS tratam de revisões de literatura (RL), contudo nenhum deles é apresentado na forma de RS. Assim, a fim de traçar um paralelo entre a presente RS e essas RLs, foi realizada uma análise livre sobre quais seriam os estudos primários abordados em cada revisão, e assim foi possível inferir o período abordado por elas. Essa análise partiu do pressuposto que os estudos primários estão principalmente descritos nas seções que tratam do tema abordado pela RL, excluindo as citações a estudos que fornecem a base teórica para entendimento do assunto. Por exemplo, em [17], a ênfase é dada em como aplicações de classificação de textos podem ser usadas em SRbC, então foram considerados como estudo primários aqueles que tratavam principalmente de aplicações de SRbC ou classificação de textos, e foram desconsiderados os trabalhos que davam base para entendimento de estratégias ou métodos.

O resultado da análise livre está na Tabela 1. Nela, a informação sobre número de estudos é apresentada de forma segmentada de acordo com o tema da RL, por exemplo, a revisão [14] apresenta 7 estudos no contexto de SRbC e 10 estudos no contexto de classificação de textos. Nessa tabela também são apresentados os dados referentes a esta RS, para fins de comparação. Os trabalhos relacionados são todos mais abrangentes que a presente RS em termos de tema. Nenhum deles é voltado para o estudo da aplicação de técnicas de AM em uma área específica em SR. Assim, um diferencial da presente RS é o estudo mais detalhado de uma linha de pesquisa dentro de SR, apresentando-se portanto como uma revisão menos abrangente, porém de maior profundidade.

A análise livre também permitiu observar que os estudos primários mais antigos abordados nas RL foram publicados na década de 1990, com exceção de [17] que inclui estudos sobre classificação de textos. Os autores de tais RL, portanto, tinham uma tendência para a apresentação de um quadro histórico dentro dos temas abordados. Diferentemente, a presente RS objetiva fornecer uma visão sobre as preocupações mais recentes existentes na área do tema abordado. A fim de apresentar os trabalhos relacionados com mais detalhes, segue aqui um breve resumo de cada um:

- [1] tem o objetivo de apresentar técnicas que são usadas em cada um dos tipos de SR em geral, assim como as estratégias para contornar as desvantagens de cada tipo de SR. Esta RL também propõem possíveis extensões e linhas para pesquisas futuras em SR.
- [13] tem o objetivo de investigar SRs no domínio de notícias, mostrando os principais problemas relacionados a este domínio. Os autores focam principalmente no problema de escalabilidade. Além disso, propõem estratégias para modelagem de perfil e como relacionar notícias com perfis. Também são realizados experimentos com as técnicas e estratégias estudadas.
- [4] tem como objetivo principal apresentar os SR no

domínio de notícias, resumizando os tipos de SR e as técnicas e estratégias neles utilizadas. O trabalho apresenta uma revisão dos tipos de SR apontando suas vantagens e desvantagens.

- [14] apresenta todos os passos referentes ao desenvolvimento de um SRbC, uma revisão sobre as técnicas utilizadas nos trabalhos, segmentando-as quanto à representação do conteúdo e aprendizado do perfil, e direções futuras para a evolução dos SRbC.
- [17] tem como objetivo principal investigar os algoritmos que são usados em SRbC, e que poderiam ser usados, como os algoritmos comumente usados na área de classificação de textos. O trabalho também discute formas para representar o conteúdo, o perfil do usuário e limitações em SRbC.

4. METODOLOGIA

Uma Revisão Sistemática (RS) permite identificar, interpretar e avaliar a pesquisa relevante para um tópico em particular [11]. Nessa estratégia, quando as contribuições individuais analisadas são chamadas de estudos primários, a RS se constitui como um estudo secundário. A diferença entre uma RS e uma revisão de literatura simples é que a primeira fornece condições de reprodutibilidade, e de avaliação do escopo abrangido na revisão e da qualidade dos estudos nela analisados. Além disso, a sistematização adotada para a elaboração de uma RS fornece condições para maximizar a possibilidade de recuperar um conjunto completo de dados e minimizar a possibilidade de um viés.

A presente RS foi conduzida com bases nas diretrizes propostas por [11], e é constituída das seguintes fases: (i) planejamento da RS; (ii) condução da RS; e (iii) resultados da RS. Nesta seção são detalhados os passos seguidos nas fases (i) e (ii). Os resultados e análises referentes à terceira fase são apresentados nas próximas seções (Seções 5 e 6).

4.1 Planejamento da Revisão Sistemática

O planejamento de uma RS inclui a identificação da justificativa para a revisão, as especificações das questões de pesquisa que se pretende responder e o desenvolvimento de um protocolo para escolha dos estudos que serão incluídos na revisão. Nesta seção, estes passos são explorados.

4.1.1 Justificativa para a Revisão Sistemática

Até onde foi possível identificar (Seção 3), não existem RS publicadas na área de AM aplicada a SRbC textual. Estratégias de AM possuem um alto potencial para fornecer conhecimento importante e útil para SR, e vêm sendo utilizadas para tal. Contudo, não foi encontrado um estudo que apresentasse os avanços que já foram obtidos e as questões em aberto em relação à aplicação de aprendizado supervisionado ou não-supervisionado nesse tipo de sistema.

SR têm sido objeto de interesse na comunidade científica e da indústria. Nos últimos anos, grandes empresas têm envidado esforços para que soluções cada vez melhores sejam obtidas em termos de recomendação. Porém, ainda há problemas em aberto como otimização da serendipidade, *cold-start* de item e de usuário e tratamento de informação multi-mídia, nos quais técnicas de AM podem atuar com sucesso. Diante deste cenário, esta RS se justifica como um meio de, sistematicamente, organizar e analisar os mais novos resultados obtidos na área de SR a partir do uso de estratégias

Tabela 1: Comparação de RLs relacionadas à presente RS

Ref.	Ano de publicação	Tema	Período	No. de Estudos Primários
[1]	2005	SR em geral	1993-2004	58
[4]	2010	SR, recomendação de notícias	1994-2007	10, 7
[13]	2011	Recomendação de notícias	1994-2011	20
[14]	2013	SRbC	1997-2008	44
[17]	2007	SRbC, classificação de texto	1961-2002	7, 10
Esta RS	2015	SRbC textual com AM	2012-2014	13

de AM. A RS apresentada aqui diz respeito aos estudos publicados nos três últimos anos (2012, 2013 e 2014).

4.1.2 Questões de pesquisa

Diante das justificativas delineadas na Seção 4.1.1 para elaboração desta RS, e seguindo o disposto em [11], um conjunto de questões de pesquisa foi elaborado.

- Q1** Em quais módulos da construção de um SRbC textual se têm empregado técnicas de AM? *a.* Qual avaliação se faz dos resultados obtidos nessas iniciativas?
- Q2** Quais tipos de AI em AM são aplicadas especificamente em SRbC textual? *a.* Quais são as vantagens e desvantagens de usar AI em SRbC textual?
- Q3** Como informações externas ao conteúdo textual sob recomendação podem ser usadas pelas técnicas de AM para melhorar os resultados do sistema?

4.1.3 Protocolo da Revisão Sistemática

O protocolo dessa RS envolveu: (i) escolha de fontes de dados e estratégias de busca; (ii) definição de estratégias para seleção de estudos primários e avaliação de sua qualidade; e (iii) método de extração de dados dos estudos selecionados.

i. Fontes de dados e estratégias de busca – As fontes de dados escolhidas para serem usadas nesta RS foram: *Scopus*, *ISI Web of Science*, *IEEE Xplore*, *ACM Digital Library* e *SpringerLink*. Cinco fontes de dados foram usadas com o intuito de maximizar o número de estudos candidatos recuperados. Essas fontes de dados foram escolhidas ou por serem temáticas (indexam veículos de disseminação científica da área de Ciência da Computação, Engenharia e áreas correlatas) ou por apresentarem mecanismos eficientes de filtragem para obtenção de estudos de áreas interessantes para o objetivo desta RS. Além disso, são notadamente fontes de dados bastante conhecidas e utilizadas pelos pesquisadores potencialmente interessados no conteúdo desta RS.

A estratégia de busca foi estabelecida a partir de uma *string* genérica (Tabela 2) a ser aplicada à cada uma das fontes de dados. A *string* é uma combinação de palavras-chave que pretende maximizar a abrangência de recuperação e trazer estudos úteis para a composição das respostas para as questões de pesquisa. Assim, a *string* foi composta por termos relacionados apenas à área de aplicação objetivada nesta RS (Sistemas de Recomendação baseados em Conteúdo textual). Já os aspectos referentes à área de AM foram tratados nos critérios de inclusão, uma vez que inúmeros termos deveriam ser usados para abranger a área de AM, pois é frequente que a área de AM esteja presente em um artigo representada apenas pelo nome de um algoritmo.

ii. Estratégias de seleção de estudos primários e de avaliação da qualidade – Para a seleção dos estudos primários a serem analisados foram estabelecidos critérios de inclusão. Um estudo primário recuperado a partir da aplicação da *string* de busca foi incluído na RS se ele atendeu

Tabela 2: Expressão regular para a *string* de busca genérica aplicada aos mecanismos de busca considerando os campos de indexação: título do estudo, resumo, palavras-chave.

("content [-]0,1 based filter*" OR "content [-]0,1 based recommend*") AND (text* OR news OR article[s]0,1 OR paper[s]0,1 OR book[s]0,1)

a todos os critérios de inclusão definidos. Tais critérios têm o objetivo de garantir que os estudos incluídos possuem um nível de qualidade mínimo (CI-1), são primários (CI-2), são acessíveis (CI-3 e CI-4) e são de fato pertinentes ao escopo da RS (CI-5 a CI-7). Os critérios de inclusão aplicados foram:

- **CI-1:** o registro de dados identificado se refere a um estudo científico, publicado através de uma revisão por pares (o registro de dados não se refere a relatórios técnicos, livros, capítulos de livros, prefácios de anais ou editoriais de periódicos, pôsteres e *position papers*);
- **CI-2:** o trabalho é um estudo primário (não se trata de estudo secundário: revisão da literatura ou RS);
- **CI-3:** o trabalho está disponível na *web*;
- **CI-4:** o trabalho é apresentado na língua inglesa;
- **CI-5:** o trabalho é diretamente relacionado com as áreas de Ciência da Computação ou Sistemas de Informação (i.e., o trabalho não é relacionado unicamente com áreas como Marketing ou Ciências Sociais);
- **CI-6:** o trabalho trata, explicitamente, SR para conteúdo textual;
- **CI-7:** o trabalho faz referência explícita ao uso de técnicas de AM para modelagem e/ou implementação de algum módulo do SR.

Os critérios de inclusão foram avaliados a partir da análise dos metadados dos artigos, do título do artigo, do resumo e do texto na íntegra. Além disso, a abordagem *Tollgate* foi aplicada [2], conforme detalhado na Seção 4.1.4. Na abordagem *Tollgate*, diferentes pesquisadores aplicam os critérios de inclusão, e a decisão final sobre a inclusão de um artigo nos resultados da RS é tomada por voto majoritário.

iii. Método de extração de dados dos estudos selecionados – A extração de dados se deu a partir da leitura, na íntegra, dos artigos selecionados para inclusão na RS. Para guiar a extração dos dados foi criado um formulário para anotação de: informações básicas como título, autores, veículo de publicação e ano de publicação; informações sobre o conteúdo do estudo em relação à área de AM e à área de SRs. Na Tabela 3 são listadas as informações sobre conteúdo anotadas nos formulários.

4.1.4 Condução da Revisão Sistemática

Esta seção apresenta a condução da revisão, a qual é dividida em duas etapas: (i) identificação e seleção dos estudos

Tabela 3: Informações sobre conteúdo extraídas dos estudos analisados nesta RS

Tópico	Descrição
Conjunto de Dados	Esse atributo contribui para o entendimento do contexto SRbC textual e dos experimentos realizados.
Análise e Repres. do Conteúdo	Verificação sobre se a análise e repres. do conteúdo foi tratada e se técnicas de AM foram aplicadas nesse tratamento.
Representação textual e pré-processamento	Para análise deste tópico foi verificado como se deu a transformação do conteúdo textual (não-estruturado) em uma representação estruturada.
Estratégia e técnicas de AM	Para esse tópico foi identificada qual o tipo de AI (supervisionada, semi-supervisionada ou não-supervisionada) usado no módulo analisador de conteúdo (Figura 1), e qual técnica foi usada para a implementação do aprendizado.
Personalização	Verificação sobre se a fase de personalização foi tratada e se técnicas de AM foram aplicadas nesse tratamento.
Representação do perfil	Para análise deste tópico foi verificado como se deu a representação do perfil do usuário.
Estratégia e técnicas de AM	Neste quesito foi identificada qual o tipo de AI (supervisionada, semi-supervisionada ou não-supervisionada) usado no Aprendizado do Perfil (Figura 1), e qual técnica foi usada para a implementação do aprendizado.
Recomendações	Verificação sobre se a fase de recomendação foi tratada e se técnicas de AM foram aplicadas nesse tratamento.
Estratégia e técnicas de AM	Neste tópico foi identificado qual o tipo de AI (supervisionada, semi-supervisionada ou não-supervisionada) usado no Componente de Filtragem (Figura 1), e qual técnica foi usada para a implementação do aprendizado.

primários e (ii) extração e síntese dos dados desses estudos.

i. Identificação e seleção dos estudos primários – Para a identificação dos estudos primários, a *string* de busca genérica foi adaptada e executada em cada uma das fontes de dados. A aplicação foi realizada no período de agosto a setembro de 2014. As bases de dados retornaram 97 estudos (para o cômputo desse número, artigos indexados em mais de uma das bases de dados foram contabilizados apenas uma vez) publicados nos anos de 2012, 2013 e 2014. Os critérios de inclusão foram aplicados seguindo os passos: as listas de registros retornados das buscas foram avaliadas para aplicação do critério CI-1; os resumos dos registros restantes foram avaliados e os critérios CI-2 e CI-5 ao CI-7 foram aplicados; os estudos referentes aos registros restantes foram avaliados segundo os critérios CI-3 e CI-4; os estudos restantes foram lidos na íntegra para extração de dados e novamente os critérios CI-6 ao CI-7 foram avaliados. Os critérios CI-1 ao CI-5 foram objeto de atenção de um pesquisador envolvido na condução da RS. Já a primeira aplicação dos critérios CI-6 ao CI-7 foi realizada por três pesquisadores usando uma abordagem *Tollgate*. Ao fim do consenso na abordagem *Tollgate*, a seleção final de artigos a serem incluídos na RS foi realizada. A segunda aplicação dos critérios CI-6 ao CI-7 foi realizada por um pesquisador.

ii. Extração e síntese de dados – A extração e síntese seguiram o preenchimento de formulários; os resultados estão consolidados e discutidos nas seções 5 e 6.

5. RESULTADOS

Nesta seção são apresentados os resultados da condução do protocolo. A aplicação dos critérios de inclusão resultou na seleção de 13 estudos (13% dos resultados recuperados), sendo que 6 foram publicados em 2012, 5 em 2013 e 2 em 2014. Embora esse resultado indique um uso ainda pequeno de AM em SRbC textual, é preciso considerar a seleção dos artigos de forma mais refinada: 51 estudos (52%) foram considerados atender ao critério CI-6, e 26 (27%) ao critério CI-7. Nesse universo reduzido, a área de AM se mostra um pouco mais presente nas soluções para SRbC textual.

Análises gerais sobre o conjunto de estudos selecionados são apresentadas nas Tabelas 4, 5 e 6. Na primeira análise (Tabela 4), os estudos foram classificados quanto à sua relação com as questões de pesquisa, indicando quais estudos contribuíram para a formulação das respostas. A segunda análise (Tabela 5) teve o objetivo de organizar os 13 estudos em relação aos aspectos: em qual módulo de um SRbC a técnica de AM é aplicada; e que tipo de aprendizado é

considerado para propor as soluções para os problemas encontrados na concepção desses módulos. Essa análise mostrou uma concentração de estudos que aplicam aprendizado supervisionado no módulo de aprendizado do perfil (8 estudos) e aprendizado não-supervisionado no módulo de representação do conteúdo (7 estudos), sendo que desses últimos, 6 estão preocupados em tratar o problema de redução de dimensionalidade. A terceira análise (Tabela 6) apresenta uma visão sobre os domínios de recomendação sob os quais os SRbC textual construídos com AM têm sido investigados.

Tabela 4: Estudos X questões de pesquisa (QP)

QP	Estudos	#
Q1	[5, 7, 8, 9, 16, 19, 26, 27, 28, 29, 30, 31]	12
Q2	[5, 7, 8, 9, 16, 19, 23, 26, 27, 28, 29, 30, 31]	13
Q3	[5, 7, 16, 19, 23, 29, 31]	7

Tabela 5: Estudos X módulos de um SRbC textual X tipos de AI

Módulo	Aprendizagem		
	sup.	não sup.	semi sup.
Representação de conteúdo	–	[5, 7, 26, 27, 28, 30, 31]	–
Aprendizado de perfil	[7, 8, 9, 16, 19, 23, 27, 29]	–	–
Filtragem	[31]	–	–

Tabela 6: Estudos X domínio de recomendação (DR)

DR	Estudos	DR	Estudos
Notícias	[5, 7, 16, 29]	Artigos	[27, 28]
Ofertas de empregos	[8]	Especialista	[26]
Conteúdo <i>web</i>	[9]	Micro-blogs	[30]
Filmes	[19]	Lugares	[23]
<i>e-Commerce</i>	[31]		

Em um trabalho de análise mais específico, cada um dos 13 estudos foi detalhadamente analisado para identificação das soluções propostas para construção dos módulos de um SRbC (conforme Figura 1). Os resultados dessa análise estão organizados por módulo nas três próximas seções.

5.1 Análise e Representação de Conteúdo

Nos estudos analisados, os módulos de Análise e Representação de Conteúdo dos SRbC textual recebem os itens tex-

tuais candidatos à recomendação, geralmente sob uma representação não estruturada, os submetem a um pré-processamento de textos, e os direcionam a técnicas de AM que geram uma representação estruturada para os próximos módulos.

Evidentemente, a fase de pré-processamento de textos está presente nos 13 estudos analisados. De forma geral, os autores usam o Vector Space Model¹ (VSM) [22] em alguma fase do processo. Porém, outros recursos também são usados em alguns estudos para estabelecimento das dimensões, como por exemplo: uso de hiperônimos do dicionário *WordNet*, *POS tag* (*Part-of-speech tagging*) para a extração de substantivos, *n-gramas* [5]; extração de *tags* relacionadas aos itens sob recomendação [16]. Uma vez escolhidas as dimensões, é preciso atribuir valores a elas, e para isso são empregadas diferentes estratégias: a *binária* [9, 31], em que é atribuído 1 se o valor de uma dimensão ocorre no documento ou 0 se não ocorre; a *frequência de termos* (TF) [5, 7, 8, 16, 19, 23, 26, 27, 28, 29], que considera o número de vezes que o valor de uma dimensão ocorre no documento; e a clássica normalização *TF-IDF* [5, 7, 8, 9, 19, 26, 29], na qual a frequência é relativa em relação ao corpus de documentos. Também na etapa de pré-processamento, são utilizadas técnicas simples para simplificar ou reduzir a dimensionalidade dos vetores de representação, como filtro de *stopwords*, que são palavras que não adicionam informação útil ao VSM [5, 8, 9, 16, 19, 27, 31]; *stemming*, que reduz cada palavra para o seu radical [5, 8, 9, 19, 27, 31]; e filtragem de termos abaixo ou acima de um limiar de frequência [5, 8].

Ainda com o intuito de gerar um modelo de representação estruturada para uso dos próximos módulos do sistema, alguns estudos aplicam técnicas de AM não-supervisionado (veja Tabela 5). Tratam-se, na realidade, de técnicas de redução de dimensionalidade que induzem modelos de representação a partir do conjunto de itens, por meio da implementação de diferentes algoritmos: determinação de variáveis latentes [7]; *Latent Semantic Analysis* (LSA) [26], *Latent Dirichlet Allocation* (LDA) [27, 30], otimização gradiente descendente [31] e *random walks* [28]. Esses algoritmos reduzem a dimensionalidade dos vetores no VSM, criando uma representação compacta dos itens. O desafio nesse contexto é descobrir o número ideal de dimensões a serem usadas na representação. A avaliação sobre o número ideal é geralmente realizada de forma empírica, por meio da realização de testes variando o número de fatores e analisando a qualidade das recomendações geradas ao final do processamento executado pelos SRs [7, 26, 27, 30].

Em [5], os autores aplicam técnicas de agrupamento sobre a representação vetorial original gerando um modelo de representação baseado em grupos de notícias. Nessa estratégia, o algoritmo *k-means* e uma variação dele aplicada a hiperônimos do dicionário *WordNet* geram grupos de notícias que suportam o fornecimento de recomendações extras. Nesse caso, o SR gera recomendações iniciais que são complementadas por outras notícias presentes nos grupos das notícias inicialmente recomendadas. Essa proposta é interessante para minimizar o problema de *cold-start* de usuário.

SRbCs podem usar também, na construção da representação do conteúdo, informação proveniente de outras fontes de dados. Estratégias nessa linha são discutidas nos estudos [7, 30, 31]. Em [7], os autores adicionam informações sobre a categoria da notícia, influenciando no cálculo das proba-

bilidades relacionadas às variáveis latentes. Nos demais artigos, embora acrescentar atributos descritivos externos ao conteúdo influencie na qualidade das recomendações geradas ao final do processo, a maneira como as técnicas de AM são aplicadas para construção do VSM não é alterada pela presença ou ausência de tais atributos.

5.2 Personalização

O módulo de Personalização, onde ocorre o aprendizado do perfil do usuário, é implementado com técnicas de AM, tipo supervisionadas, em 8 estudos (Tabela 5). Nele, o perfil do usuário pode ser representado no mesmo espaço (VSM) que o item sob recomendação, por meio do uso de termos que ocorrem em documentos que eles acessam, como ocorre nos estudos [7, 8, 9, 16]. Contudo, alguns autores propõem algumas representações alternativas: em [29] os autores incorporam informações demográficas na tentativa de minimizar o problema de *cold-start* de usuário; em [8] informações referentes aos usuários, provenientes de redes sociais, são utilizadas; e, finalmente, em [23], o enriquecimento do perfil do usuário é realizado com a aplicação de AM, representado no referido estudo por *Hidden Markov Models* e Árvore de Decisão, para extração de informações referentes a localização do usuário, malha de transporte e locais de visitação.

Nos estudos que aplicaram AM para personalização foram construídos classificadores que dado um item (como entrada), o classifica como adequado para o usuário ou não. Usualmente um classificador é treinado para cada usuário com base no histórico de iterações entre o usuário e os itens. Experiências positivas entre usuário e item determinam a classe positiva (o item é adequado para o usuário), e experiências negativas determinam a classe negativa. Diversos algoritmos de AM supervisionado são usados nessa estratégia: modelos probabilísticos [7, 29], *Support Vector Machine* (SVM) [8, 9, 27], classificador de Rocchio [8, 9], *k-nearest neighbor* (k-NN) [19] e Regressão Logística (RL) [16].

Em [7] e [29], os autores usam os modelos probabilísticos *Expectation-Maximization* para geração de modelos generativos, e Redes Bayesianas, respectivamente, para indicar se um usuário lerá ou não uma notícia. Experimentos em [7] mostraram que a estratégia capta interesses de curto e longo prazo, usando conjunto de dados estratificados de acordo com a variável de tempo, e avaliando o resultado com uso das métricas *P@k* e *Normalized Discounted Cumulative Gain*, as quais analisam a posição em que a notícia apareceu na lista de recomendação. Ainda no contexto de notícias, em [16], os interesses do usuário são modelados com RL.

Classificadores de Rocchio e SVM são usados em [8] e [9], seguindo a estratégia usual de construção de classificadores em SRs. Em ambos os casos, experimentos avaliados sob as clássicas métricas baseadas em matriz de confusão (área sob a curva ROC em [8] e precisão, revocação e *f-score* em [9]) indicaram a superioridade da técnica SVM.

SVMs também são usados em [27] para decidir se um artigo é relevante ou não para os usuários do SR, dada a *string* de busca que o usuário entra no sistema. Ainda, [19] aplica como base para recomendação, uma tarefa de regressão que faz o uso do algoritmo k-NN para prever a nota que o usuário pode dar a um filme.

5.3 Recomendação

O módulo de Recomendação recebe a lista de recomendações gerada no módulo de Personalização e a disponibiliza

¹No VSM, as dimensões do vetor que representa o item (ou documento) são compostas por palavras nele presentes.

ao usuário, refinando-a se necessário. Apenas um estudo [31] realiza processamento de informação nesse módulo usando uma técnica de AM, a k -NN. Em tal proposta, os autores aplicam k -NN na representação do itens, sem que o perfil do usuário seja construído. Os k itens mais similares a um dado item (acesso pelo usuário) entram na lista de recomendação.

6. DISCUSSÕES

Nesta seção são apresentadas respostas para as questões de pesquisa elaboradas para essa revisão sistemática.

Q1. *Em quais módulos da construção de um SRbC textual se têm empregado técnicas de AM?* – A análise dos estudos selecionados nessa revisão sistemática permitiram concluir que, considerando a arquitetura típica de um SRbC (Figura 1), AM oferece recursos para resolução de problemas nos três módulos: de Análise e Representação de Conteúdo, de Personalização e de Recomendação. No caso do primeiro módulo, as técnicas de AM são predominantemente aplicadas com o objetivo de redução de dimensionalidade na representação dos itens (sob recomendação). No segundo módulo, Aprendizado de Perfil, as análises evidenciam que há potencial para aplicação de AM em, pelo menos, duas frentes: para enriquecimento da representação do perfil do usuário; e para aprendizado do perfil. Finalmente, no terceiro módulo, a aplicação de AM foi observada em apenas um estudo, no entanto, trata-se de fato de um módulo bastante simplificado na arquitetura. É interessante observar que a estratégia usada para aplicar AM no terceiro módulo substituiu a necessidade de implementação do segundo módulo (sem aparente prejuízo de desempenho), possibilitando a otimização da arquitetura do sistema.

Q1.a *Qual avaliação se faz dos resultados obtidos nessas iniciativas?* – Todos os estudos apresentaram experimentos cujos resultados foram avaliados por meio de medidas classicamente usadas na área de AM. Os resultados obtidos estão dentro do que se considera resultados aceitáveis na área. Além disso, os experimentos apresentados na maior parte dos artigos também fizeram avaliações empíricas, comparando os resultados das recomendações com dados históricos referentes às interações dos usuários com os itens recomendados. Tais comparações também evidenciaram a adequabilidade das técnicas de AM ou sua superioridade.

Q2. *Quais tipos de AI em AM são aplicadas especificamente em SRbC textual?* – A aplicação de AI não-supervisionados foi observada na implementação de otimizações no módulo de análise e representação de conteúdo. Já a aplicação do AI supervisionado foi observada nos outros dois módulos. Não foi observada o uso de AI semi-supervisionado. Ainda, dentro dos AI não-supervisionado, apenas um estudo utilizou técnicas de quantização do espaço (como o k -means). Naturalmente, o tipo de tarefas que parecem ser interessantes para cada um dos módulos levou a esse viés. O AI supervisionado é mais facilmente aplicado no contexto onde se tem uma base de dados histórica que pode ser rotulada, como é caso da tarefa de aprendizado de perfil. Geralmente, há informação sobre que item foi, no passado, acessado por qual usuário, o que permite supervisionar a indução de um modelo que represente essa relação.

Q2.a *Quais são as vantagens e desvantagens de usar AI em SRbC textual?* – A implementação da redução de dimensionalidade na representação dos itens levou a melhorias de

desempenho em alguns estudos [7, 26, 27, 28, 30, 31], principalmente por possibilitar a proposição de tópicos para a representação. Entretanto, as técnicas que se baseiam em *Singular Value Decomposition*, como LSA, não permite a interpretação dos fatores latentes, que representariam os tópicos. Os estudos que implementaram classificadores para o aprendizado do perfil do usuário alcançaram bom desempenho, principalmente no uso de SVM. Contudo, seguindo as abordagens propostas nos artigos, pode-se cair em problemas de escalabilidade devido à necessidade da construção de um classificador por usuário. Além disso, a indução desses classificadores geralmente exige uma grande quantidade de exemplos rotulados, o que nem sempre é factível.

Q3. *Como informações externas ao conteúdo textual sob recomendação podem ser usadas pelas técnicas de AM para melhorar os resultados do sistema?* – O uso de informações externas ao conteúdo textual foi observada em sete estudos [5, 7, 16, 19, 23, 29, 31]. Em [7] e [23], informações externas estão diretamente relacionadas à forma como as técnicas de AM são utilizadas. No primeiro caso, probabilidades usadas na construção da representação do conteúdo são influenciadas pelo uso de informações sobre as classes das notícias, e no segundo caso, as técnicas de AM são usadas para obtenção da informação externa. Ainda em [7], informação referente à data de publicação de notícias é usada para estratificar o conjunto de dados de treinamento para indução da recomendação, porém, a forma como a técnica de AM é usada não é alterada. Isso também ocorre nos demais estudos que inserem informações externas como atributos descritivos dos itens de recomendação: [5, 16] complementam a representação com informações de ontologias e de artigos do *wikipedia*, respectivamente; [19, 31] utilizam outros tipos de informações, como imagem e som. Informações externas também são usadas para complementar a representação do perfil dos usuários, [29] aplicam informações demográficas dos usuários para complementar a representação de seus perfis. Melhorias de desempenho observadas em tais estudos são obtidas porque informações externas aumentam o poder descritivo da representação, e não por estarem diretamente ligadas à maneira como as técnicas de AM são aplicadas. Não foi encontrado nenhum estudo que fizesse uso de qualquer conhecimento externo ao conteúdo textual que pudesse ser utilizado pelas técnicas de AM para obter vantagens específicas, como por exemplo, determinar número de grupos ou número de variáveis latentes, ou ainda permitir o uso de estratégias de *boosting* para o algoritmo de aprendizado.

7. CONCLUSÃO

Neste artigo foi apresentada uma Revisão Sistemática sobre a intersecção das áreas: SRbC textual e AM. Com a condução da revisão foi observado que iniciativas usando AM em SRbC textual ainda são em pequeno número. Entretanto, na análise dos estudos selecionados percebeu-se que as iniciativas na intersecção dessas áreas têm produzido resultados promissores, fornecendo uma motivação para que mais esforços de pesquisa sejam empregados nessa área. Nesse contexto, foi ainda possível perceber que há uma carência de iniciativas voltadas para estudar a adequabilidade de AI não-supervisionado na tarefa de Aprendizado do Perfil do usuário. Técnicas de aprendizado não-supervisionado podem contribuir nesta tarefa resolvendo desvantagens encontradas no uso de AI supervisionado (escalabilidade e necessi-

dade de dados rotulados ou *cold-start* de usuário), ou produzindo conhecimento inesperado em uma recomendação, minimizando o problema de serendipidade, comum em SRbC.

8. REFERÊNCIAS

- [1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. 17(6):734–749, 2005.
- [2] W. Afzal, R. Torkar, and R. Feldt. A systematic review of search-based testing for non-functional system properties. 51(6):957–976, 2009.
- [3] D. Bell, J. Guan, and Y. Bi. On combining classifier mass functions for text categorization. 17(10):1307–1319, Oct 2005.
- [4] H. L. Borges and A. C. Lorena. A survey on recommender systems for news data. In E. Szczerbicki and N. Nguyen, editors, *Smart Inf. and Knowledge Management*, volume 260 of *Studies in Comput. Int.*, pages 129–151. Springer, 2010.
- [5] C. Bouras and V. Tsogkas. Assisting cluster coherency via n-grams and clustering as a tool to deal with the new user problem. pages 1–14, 2014.
- [6] D. Cai and X. He. Manifold adaptive experimental design for text categorization. 24(4):707–719, April 2012.
- [7] S. Cleger-Tamayo, J. M. Fernández-Luna, and J. F. Huete. Top-n news recommendations in digital newspapers. 27:180–189, 2012.
- [8] M. Diaby, E. Viennet, and T. Launay. Exploration of methodologies to improve job recommender systems on social networks. 4(1), 2014.
- [9] D. Godoy. Comparing one-class classification algorithms for finding interesting resources in social bookmarking systems. In *Resource Discovery*, volume 6799 of *LNCS*, pages 88–103. Springer, 2012.
- [10] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems An Introduction*. Cambridge Univ. Press, 2011.
- [11] B. Kitchenham. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, Keele Univ., UK, 2007.
- [12] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. 42(8):30–37, Aug. 2009.
- [13] L. Li, D. Wang, S. Zhu, and T. Li. Personalized news recommendation: A review and an experimental investigation. 26(5):754–766, 2011.
- [14] P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 73–105. Springer, 2011.
- [15] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.
- [16] F. Narducci, C. Musto, G. Semeraro, P. Lops, and M. de Gemmis. Exploiting big data for enhanced representations in content-based recommender systems. 152:182–193, 2013.
- [17] M. J. Pazzani and D. Billsus. The adaptive web. chapter Content-based Recommendation Systems, pages 325–341. Springer, Berlin, Heidelberg, 2007.
- [18] S. M. Peres, T. Rocha, M. R. C. B. Bísaro, H. H., and C. Boscarioli. Tutorial sobre fuzzy-c-means e fuzzy learning vector quantizations: Abordagens híbridadas para tarefas de agrupamento e classificação. 19(1):120–163, 2012.
- [19] W. Qu, K.-S. Song, Y.-F. Zhang, S. Feng, D.-L. Wang, and G. Yu. A novel approach based on multi-view content analysis and semi-supervised enrichment for movie recommendation. 28(5):776–787, 2013.
- [20] P. Resnick and H. R. Varian. Recommender systems. 40:56–58, March 1997.
- [21] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 1–35. Springer, 2011.
- [22] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. 18(11):613–620, 1975.
- [23] S. Savage, M. Baranski, N. E. Chavez, and T. Hollerer. I’m feeling loco: A location based context aware recommendation system. In *Advances in Location-Based Services: 8th International Symposium on Location-Based Services, Vienna 2011*, LNCS. Springer, 2011.
- [24] F. Sebastiani. Machine learning in automated text categorization. 34(1):1–47, 2002.
- [25] S. Senecal and J. Nantel. The influence of online product recommendations on consumers’ online choices. 80(2):159–169, 2004.
- [26] A. Spaeth and M. Desmarais. Combining collaborative filtering and text similarity for expert profile recommendations in social websites. In *User Modeling, Adaptation, and Personalization*, volume 7899 of *LNCS*, pages 178–189. Springer, 2013.
- [27] S. Tantanasiwong. A comparison of clustering algorithms in article recommendation system. In *Proc. of SPIE - The Int. Soc. for Optical Eng.*, volume 8349, Singapore, 2012.
- [28] Y. Wang, J. Liu, X. Dong, T. Liu, and Y. Huang. Personalized paper recommendation based on user historical behavior. In M. Zhou, G. Zhou, D. Zhao, Q. Liu, and L. Zou, editors, *Natural Language Processing and Chinese Comp.*, volume 333 of *Commun. in Comp. and Inf. Sci.*, pages 1–12. Springer, 2012.
- [29] K. F. Yeung and Y. Yang. A proactive personalized mobile news recommendation system. In *J. of Internet Services Applications (2012)*, pages 207–212. IEEE, Sept. 2012.
- [30] J. Yu, Y. Shen, and J. Xie. Mining user interest and its evolution for recommendation on the micro-blogging system. In J. Wang, H. Xiong, Y. Ishikawa, J. Xu, and J. Zhou, editors, *Web-Age Information Management*, volume 7923 of *LNCS*, pages 679–690. Springer, 2013.
- [31] Q. Zhu, Z. Li, H. Wang, Y. Yang, and M.-L. Shyu. Multimodal sparse linear integration for content-based item recommendation. In *Multimedia (ISM), 2013 IEEE Int. Symp. on*, pages 187–194, Dec 2013.

Projeto e implementação de uma rede neural artificial para detecção do mal-posicionamento rotacional de dedos em dispositivos de captura de impressões digitais multivista sem toque

Alternative Title: Design and implementation of an artificial neural network applied to finger bad-positioning detection on touchless multiview fingerprints devices

Cauê Zaghetto
Prog. de Pós-graduação em
Sistemas Mecatrônicos
Universidade de Brasília
Brasília, Brasil
cauezag@aluno.unb.br

Luiz Henrique M. Aguiar
Prog. de Pós-graduação em
Sistemas Mecatrônicos
Universidade de Brasília
Brasília, Brasil
luizhenrique@aluno.unb.br

Alexandre Zaghetto
Depart. de Ciência da
Computação
Universidade de Brasília
Brasília, Brasil
alexandre@cic.unb.br

Flávio de Barros Vidal
Depart. de Ciência da
Computação
Universidade de Brasília
Brasília, Brasil
fbvidal@unb.br

RESUMO

Este artigo apresenta um método baseado em Redes neurais artificiais que avalia o mal-posicionamento dos dedos, em função da rotação, em dispositivos de aquisição de impressões digitais multivista sem toque. O objetivo é determinar se o dedo está rotacionado ou não, uma vez que o bom posicionamento do dedo é mandatório para garantir altas taxas de correspondência de impressões digitais. Um conjunto de teste de 9000 imagens foi usado para treinar, validar e testar um classificador baseado em redes neurais artificiais multicamadas. Até o momento, não existe um método definitivo que abordou o problema de qualidade de impressões digitais em scanners que utilizem a tecnologia multivista sem toque e a detecção da rotação do dedo apresentada neste artigo é um dos passos que devem ser levados em conta se um futuro método automático para avaliação da qualidade de impressões digitais for considerado. Os resultados médios, mostram que: (a) o classificador identifica corretamente o mal-posicionamento em aproximadamente 94% dos casos; e (b) quando o mal-posicionamento é detectado, o ângulo de rotação é corretamente estimado em 90% dos casos.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

Palavras-Chave

impressão digital, multivista, sem toque, avaliação de qualidade.

ABSTRACT

This paper presents a method based on Artificial Neural Network that evaluates the rotational bad-positioning of fingers on touchless multiview fingerprinting devices. The objective is to determine whether the finger is rotated or not, since a proper positioning of the finger is mandatory for high fingerprint matching rates. A test set of 9000 acquired images has being used to train, validate and test the proposed multi-layer Artificial Neural Network classifier. To our knowledge, there is no definitive method that addressed the problem of fingerprint quality on touchless multiview scanners. The proposed finger rotation detection here presented is one of the steps that must be taken into account if a future automatic image quality assessment method is to be considered. Average results show that: (a) our classifier correctly identifies bad-positioning in approximately 94% of cases; and (b) if bad-positioning is detected, the rotation angle is correctly estimated in 90% evaluations.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

General Terms

Fingerprint quality assessment

Keywords

fingerprint, multiview, touchless, quality assessment.

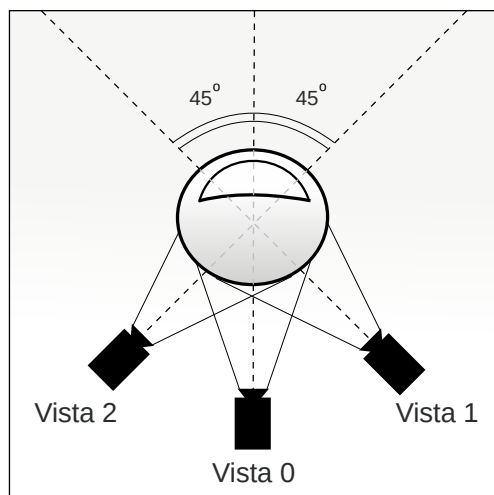


Figura 1: Descrição de um sistema de captura para impressão digital multivista sem toque, utilizando três câmeras (uma central e duas laterais) para captura da área central e lateral da impressão digital.

1. INTRODUÇÃO

A autenticação biométrica, ou simplesmente biometria, pode ser definida como a verificação ou reconhecimento automático de identidade de um indivíduo com base nas características fisiológicas e comportamentais [19]. Impressão digital, geometria da mão, voz, íris, face, escrita à mão e forma/tempo de digitação no teclado são exemplos de tais características. Diferentes sistemas biométricos requerem tecnologias específicas, dependendo do aspecto fisiológico (ou comportamental) que está sendo usado [18]. No entanto, a biometria baseada em impressão digital é hoje o sistema mais usado e reconhecido pela indústria e academia. Se comparado a outros parâmetros biométricos, a impressão digital é a mais popular e amplamente utilizado em todo o mundo. Uma grande parte de suas aplicações está associada à identificação civil e investigação criminal. Praticamente, todos os departamentos policiais [10] possuem um Sistema Automático de Identificação de Impressões Digitais (AFIS) [9]. Apesar de sua maturidade e ampla divulgação, acredita-se que os algoritmos de aquisição de impressão digital, processamento de imagens e de correspondência não estão no seu potencial pleno e definitivo [16].

Dessa forma, uma análise mais cuidadosa mostra que há muito espaço para melhorias não só para aprimorar o estado da arte dos algoritmos e processamento, mas também para mostrar novas e melhores formas de realizar a biometria. Nós descrevemos brevemente uma nova, mas já conhecida, tecnologia de aquisição de impressões digitais multivista sem toque, ilustrada na Figura 1. Esta resolve muitos problemas de scanners baseados em toque, mas é suscetível a novos. O problema mais fundamental desta tecnologia reside no fato de que é preciso garantir o correto posicionamento do dedo em relação à câmara de aquisição. Este problema surge a partir do fato de que no processo de aquisição utilizando um dispositivo sem uma superfície de toque, os dedos estão sujeitos a mais graus de liberdade. Até onde se sabe, não existe um método definitivo, que aborda o problema da qualidade da impressão digital para scanners multivistas sem toque.

Portanto, neste artigo apresentamos um algoritmo que indica se o dedo foi colocado corretamente ou não, devido à possibilidade de mal-posicionamento rotacional do mesmo. Este é o primeiro passo para a definição de um método de avaliação de qualidade mais completo, que pode orientar o processo de aquisição de impressões digitais sem toque no futuro. Desta feita, as informações do processo identificação de impressões digitais e metodologia proposta para se alcançar os resultados são descritos na Seção 2 e 3, respectivamente. Os resultados são apresentados e discutidos na Seção 4. Finalizando na Seção 5, em que as conclusões e os trabalhos futuros são apresentados.

2. IDENTIFICAÇÃO DE IMPRESSÕES DIGITAIS

Em uma forma simplificada, o processo de identificação de impressões digitais pode ser subdividido em quatro sub-processos: (1) Captura da imagem; (2) Pré-processamento; (3) Extração de Características; e; (4) Correspondência das características. Entre as diversas classificações que podem ser realizadas na identificação de impressões digitais, a maioria faz do uso de técnicas de capturas da imagem com a necessidade do toque entre o dedo do usuário e o dispositivo de captura utilizado. Atualmente o processo de captura é dividido de acordo com o tipo de tecnologia: com ou sem toque do dedo (i. e. *touchless*) do usuário no dispositivo. Se por um lado, sistemas em que o toque no sensor se faz necessário, já alcançaram um nível de maturidade, por outro, sistemas que utilizam a tecnologia sem toque no sensor ainda estão nos seus primórdios, permitindo ainda diversos avanços no processo de identificação biométrica, a partir da impressão digital.

2.1 Impressões digitais com toque

Atualmente diversas aplicações nas áreas forenses, civis e comerciais [1] fazem uso de impressões digitais. Nestes sistemas, a autenticação a partir do uso de imagens digitais das impressões digitais, permitem uma descrição real dos dedos, aumentando a confiabilidade do sistema. Desta feita, o processo de aquisição se torna um passo crucial do sistema, interferindo diretamente no desempenho geral do processo de identificação. A maioria dos sistemas atuais empregam tecnologias baseadas no toque do dedo do usuário com o sensor de captura, demandando ao usuário pressionar o dedo contra sua superfície.

Dentre os diversos problemas relacionados ao uso deste tipo de tecnologia, podemos citar as distorções e inconsistências causadas, durante o processo de captura, pela elasticidade da pele. A qualidade da impressão digital obtida pode ser também influenciada pela sujeira, suor, secura excessiva, umidade do ar, temperatura e impressões digitais danificadas (cicatrizes). Também as variações no posicionamento do dedo em cada captura, pode ocasionar em amostras diferentes de imagens capturadas para o mesmo dedo. Em algumas situações, estas desvantagens apresentadas em [8], impõem a necessidade de operadores extremamente especializados e treinados, requerendo maior quantidade de tentativas, e proporcionalmente elevando o tempo necessário para um processo de captura para todos os dedos. Por exemplo, para o cadastramento de grandes populações, como o cadastramento biométrico eleitoral realizado em 2013 e 2014 no Brasil, demandam de muito tempo para que as impressões

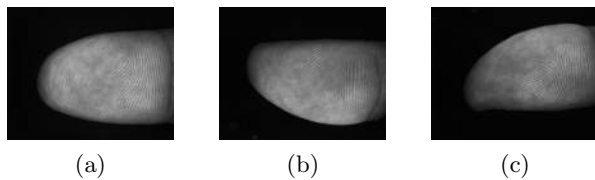


Figura 2: Exemplo de impressões digitais capturadas (1024×1280 pixels, 8 bits/pixels) capturadas por um sistema sem toque com três câmeras: (a) Vista 0; (b) Vista 1; and (c) Vista 2.

digitais capturadas tenham a qualidade necessária para o processo de identificação.

Com o passar dos anos, diversos algoritmos foram propostos para compensar essas limitações das tecnologias baseadas em toque, sendo que em alguns, estas melhorias geram gargalos para melhorias futuras na qualidade da imagem digital proveniente destas impressões digitais capturadas.

2.2 Impressões digitais sem toque

Atualmente, diversas soluções baseadas na captura de impressões digitais sem toque tem sido desenvolvidas. Todas estas para colaborar no processo de melhoria dos problemas intrínsecos dos sistemas, em que utilizam o toque no sensor durante o processo de captura [18, 11, 12, 13]. A idéia básica de sistemas sem toque é atacar o problema de qualidade da amostra em seu nível fundamental: durante o processo de captura. No caso da utilização destes sistemas, não existe a necessidade do usuário em pressionar o dedo contra o sensor, evitando os problemas listados anteriormente, e ainda, não dependendo da utilização de algoritmos para a melhoria da qualidade da amostra capturada para compensar os problemas causados pela elasticidade da pele humana.

Dentre as soluções tecnológicas disponíveis, a empresa suíça TBS¹ desenvolveu um dispositivo [17] que utiliza uma metodologia promissora. Este dispositivo combina técnicas baseada na captura da reflexão da imagem do dedo a ser capturado, com um sistema de três câmeras sincronizadas em posicionamentos diferentes (múltiplas vistas). Uma câmera é posicionada para capturar a imagem proveniente do posicionamento normal ao dedo, em que a parte principal da impressão digital está localizada (deltas e bifurcações), a partir de uma câmera central. As outras duas câmeras são posicionadas a 45° , sendo uma no sentido horário (SH) e a outra no sentido anti-horário (SAH). A partir desta abordagem, é garantida a captura sobreposta de todas as áreas da impressão digital capturada, como apresentado na Figura 1. A imagem final é obtida a partir do mosaico formado das três imagens utilizando algoritmos de alinhamento [14]. Nas Figuras 1 e 2, são apresentadas uma representação esquemática de um sistema de captura sem toque utilizando múltiplas vistas e imagens capturadas por este tipo de dispositivo, respectivamente.

2.3 Avaliação da qualidade de impressões digitais

O desempenho e interoperabilidade dos sistemas biométricos dependem da qualidade do sinal adquirido [2]. Entretanto, uma metodologia para a estimação da qualidade

¹<http://www.tbsinc.com/>

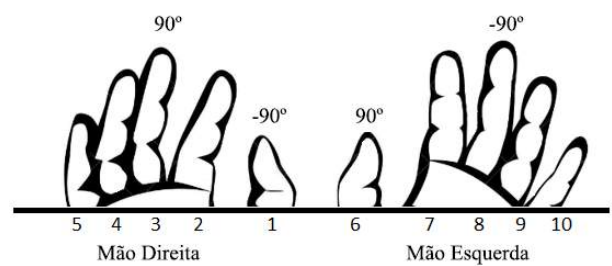


Figura 3: Identificação dos dedos e possíveis direções de rotações.

das amostras biométricas coletadas são de fundamental importância, principalmente para se obter elevados índices de acerto.

Existe uma grande variedade de tecnologias que realizam a captura da impressão digital. Mesmo com os avanços e aprimoramentos muitos desses sistemas de capturas falham no processo de obtenção de imagens com boa qualidade. Alguns dispositivos de captura podem ser melhores do que outros, sendo até imunes a fatores que afetam diretamente a qualidade da imagem produzida, mas o problema em se obter imagens com boa qualidade ainda permitem avanços. Em trabalhos recentes, vários problemas que afetam a qualidade da impressão digital (como por exemplo toques com o sensor inconsistentes, não-uniformes e irreprodutíveis [15]) propõem técnicas que resolvem estes problemas ([6, 4, 5]). Sendo assim, a análise da avaliação da qualidade são originárias dos mesmos princípios dos sistemas baseados na captura com o toque do dedo do usuário no sensor, estas propostas pelo National Institute of Standards and Technology (NIST) [20], fazendo necessária na utilização de sistemas sem toque se adequem aos mesmos níveis de exigências do sistema com toque.

3. METODOLOGIA PROPOSTA

Uma vez que o objetivo é avaliar o mal-posicionamento dos dedos em dispositivos de impressão digital sem toque, o primeiro passo é criar um conjunto de testes. As amostras foram adquiridas a partir de 20 voluntários que cedaram suas impressões digitais capturadas por um dispositivo com características gerais semelhantes às descritas na Figura 1. Amostras provenientes de todos os dedos de ambas as mãos foram coletadas. Aqui, a amostra é definida como um conjunto de três imagens (vista 0, 1 e 2) de cada dedo, como mostrado na Figura 2. Todos os dedos (10 dedos) geram 15 amostras, o que perfaz um total de 45 imagens por dedo. Portanto, o nosso conjunto de teste é composto por $20 \times 10 \times 15 \times 3 = 9000$ imagens. As 15 amostras de cada dedo são distribuídas da seguinte forma:

- 5 amostras não rotacionadas;
- 5 amostras rotacionadas em 180° ; e
- 5 amostras rotacionadas em 90° sentido horário (SH) ou sentido anti-horário (SAH), dependendo da mão (esquerda ou direita).

Quatro dedos da mão direita (2, 3, 4 e 5) e um dedo da mão esquerda (6) foram rotacionados no sentido horário. Quatro dedos da mão esquerda (7, 8, 9 e 10) e um

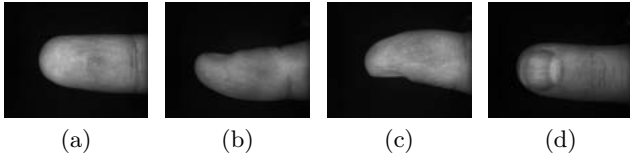


Figura 4: Efeito de diferentes ângulos de rotação na Vista 0: (a) não-rotacionado; (b) 90° SH; (c) 90° SAH; and (d) rotacionado em 180°.

dedo da mão direita (1) foram rotacionados no sentido anti-horário. Os polegares foram rotacionados no sentido oposto, considerando os outros dedos da mesma mão, devido a razões ergonômicas. A Figura 3 identifica cada dedo e os respectivos sentidos de rotação. O efeito dos diferentes sentidos de rotação sobre a vista 0 é exemplificada na Figura 4. Consequentemente, o conjunto de testes é caracterizado por 500 amostras de dedos rotacionados em 90° sentido horário, 500 amostras de dedos rotacionados em 90° sentido anti-horário, 1.000 amostras de dedos rotacionados em 180° e 1000 amostras de dedos não rotacionados, o que perfaz um total de 3.000 amostras.

O próximo passo é definir as métricas que serão utilizados como parâmetros para o classificador. A primeira métrica é a média da área útil em cada vista (m_0, m_1 e m_2). A segunda métrica é a variância destas mesmas áreas úteis (σ_1, σ_2 e σ_3). Nós assumimos que a luminância média e a textura irão variar dependendo da porção do dedo a ser fotografado. A unha produz, sem dúvida, diferentes padrões de reflexão em comparação à parte do dedo que contém o núcleo e o delta.

Para calcular os valores destas duas métricas, primeiro uma área útil deve ser delimitada. Inicialmente, aplicamos um processo de binarização com o método de Otsu [21]. Uma vez que o fundo é muito mais escuro do que o primeiro plano, o processo retorna uma segmentação do dedo muito precisa. Em seguida, o elemento de imagem branco mais à esquerda é considerado como a ponta de um dedo. Considerando-se que o comprimento do dedo L em pixels pode ser estimado como o número de colunas entre a ponta do dedo coluna f_c e a última coluna da imagem, a região de interesse (ROI) é definida como todos os pixels brancos que encontram-se entre f_c e $f_c + L/2$. A única garantia é uma área significativa do dedo estar representada dentro desta região. Pelo menos a unha deve ser incluída. A Figura 5 ilustra como a região de interesse é capturada. Além das métricas descritas anteriormente, as razões entre a largura dos dedos projetadas em cada vista também são consideradas. Como mostrado na Figura 6 as razões entre as larguras das vistas 1 (w_1) e 2 (w_2) em relação à vista 0 (w_0) podem ser utilizadas como um parâmetro discriminativo.

Uma vez que as métricas são apresentadas, o vetor de características $\mathbf{X}$, que vai ser entrada para o classificador, pode ser definido. Este vetor é composto por 6 elementos, de x_1 a x_6 , descrito como se segue:

$$\mathbf{X} = [m_1/m_0 \quad m_2/m_0 \quad \sigma_1/\sigma_0 \quad \sigma_2/\sigma_0 \quad w_1/w_0 \quad w_2/w_0] \quad (1)$$

Com intuito de detectar quando o dedo está, ou não, de-

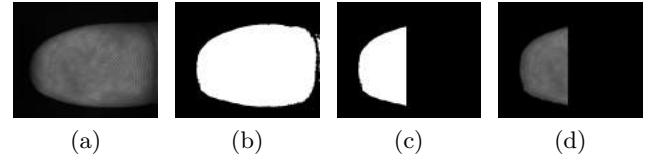


Figura 5: Recuperação da região de interesse (Vista 0): (a) Imagem original; (b) Resultado da binarização; (c) Seleção da Máscara; e (d) área selecionada para o cálculo da média e variância. O mesmo procedimento é aplicado para as vistas 1 e 2.

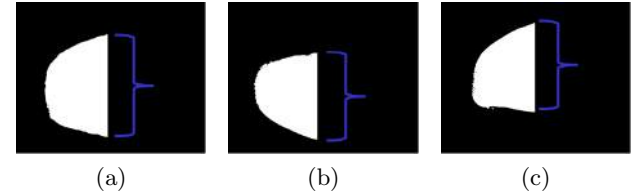


Figura 6: Dedo projetado em diferentes vistas de uma amostra: (a) Projeção da Vista 0; (b) Projeção da Vista 1; e (c) Projeção da Vista 2. Razões entre as larguras das Vistas 1 e 2 em relação à Vista 0 também podem ser usadas como um parâmetro discriminativo no vetor de características.

vidamente posicionado durante a aquisição biométrica, uma Rede Neural Artificial é proposta. O classificador é um *Perceptron* Multicamada [7] treinado com o algoritmo de *Levenberg-Marquardt* [22] que tem quatro camadas: (a) uma camada de entrada, onde nenhum cálculo é realizado; (b) duas camadas intermediárias; e (c) uma camada de saída. A primeira camada simplesmente insere vetor $\mathbf{X}$ para a rede neural proposta. A primeira e a segunda camadas ocultas têm 20 e 5 neurônios, respectivamente. A camada de saída tem um único neurônio. Exceto para a camada de entrada, todos os neurônios têm a função de ativação Tangente Hiperbólica. Sabe-se que as redes neurais sem realimentação (*feedforward*) com apenas uma camada escondida podem aproximar satisfatoriamente qualquer função com um número finito de descontinuidades arbitrárias [3]. No entanto, após testes exaustivos, observamos em nossos experimentos que melhores desempenhos foram alcançados com duas camadas ocultas. A arquitetura proposta é resumida na Figura 7.

Em nossos experimentos, dois cenários diferentes foram considerados. Em um primeiro cenário, o classificador apenas indica se existe um mal-posicionamento rotacional do dedo ou não. Em um segundo, caso o classificador tenha detectado mal-posicionamento rotacional, também detecta o ângulo de rotação (90° horário, 90° sentido anti-horário e 180°).

$$\alpha(T) = \begin{cases} 180^\circ, & \text{se } -1 \leq T \leq -0.65 \\ 90^\circ \text{ SAH}, & \text{se } -0.65 < T \leq 0.05 \\ 90^\circ \text{ SH}, & \text{se } 0.05 < T \leq 0.70 \\ 0^\circ \text{ (Não-rotacionado)}, & \text{se } 0.70 < T \leq 1 \end{cases} \quad (2)$$

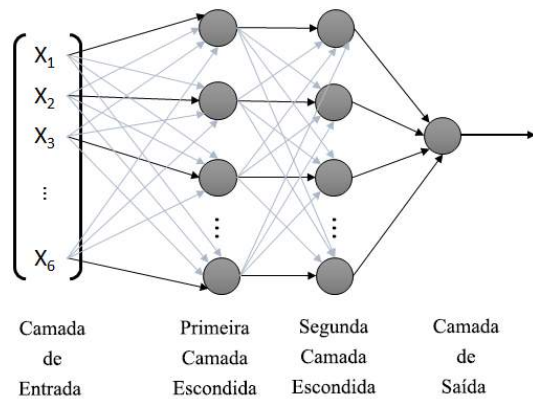


Figura 7: Arquitetura proposta para a rede neural artificial utilizada: camada de entrada com 6 neurônios, de x_1 a x_6 ; primeira camada oculta com 20 neurônios; segunda camada oculta com 5 neurônios; e camada de saída com 1 neurônio. Todos os neurônios possuem como função de ativação a tangente hiperbólica.

No primeiro cenário, o neurônio da camada de saída retorna $T = -1$ (no caso de rotação) ou $T = 1$ (no caso de correto posicionamento). Quanto ao segundo cenário, os valores de saída são quantificados em um determinado ângulo de rotação α , de acordo com Eq. 2.

O primeiro conjunto de testes usado em nossos experimentos é composto por todas as amostras disponíveis: 1000 dedos não rotacionados (20 indivíduos, 10 dedos diferentes por indivíduo, cinco amostras por dedo) e 2.000 dedos rotacionados (20 indivíduos, 10 dedos diferentes por indivíduo, 10 amostras por dedo: todos os dedos rotacionados por 180° , e por 90° SH ou 90° SAH, dependendo de cada dedo). Cerca de $1/3$ de cada classe, não rodados e rodados (por 180° , 90° SH ou 90° SAH) foi utilizado para compor o conjunto de treinamento, validação e teste da rede neural. O conjunto de treinamento é usado para atualizar as sinapses e biases da rede, o conjunto de validação é usado para monitorar o processo de formação e evitar super-ajuste. O conjunto de teste é utilizado para avaliar o desempenho da rede e não é utilizado durante o treinamento.

Um caso especial também foi investigado. Uma vez que o polegar é um dedo largamente utilizado na identificação biométrica, os dois cenários descritos acima também são aplicados para este dedo apenas. Neste caso, o subconjunto de ensaio é composto por 600 amostras de polegares: 200 representando a classe de dedos não rotacionados e 400 representando a classe de dedos rotacionados. Também neste caso, cerca de $1/3$ de cada classe foi distribuída entre o treinamento, validação e conjunto de teste.

4. RESULTADOS EXPERIMENTAIS

Os testes foram divididos em quatro categorias:

- Rotacionado $\times$ Não rotacionado: detecção de rotação para todos dedos; e;
- Rotacionado $\times$ Não rotacionado: detecção de rotação dos polegares.

Tabela 1: Resultado da detecção de todos os dedos: Rotacionado $\times$ Não-rotacionado.

Real	Predição	
	Não rot.	Rot.
	Rot.	
Não rot.	86.90%	13.10%
Rot.	7.00%	93.00%

Tabela 2: Resultado da detecção do polegar: Rotacionado $\times$ Não-rotacionado.

Real	Predição	
	Não rot.	Rot.
	Rot.	
Não rot.	96.50%	3.50%
Rot.	2.75%	97.25%

- Detecção do ângulo de rotação de todos os dedos; e;
- Detecção do ângulo de rotação dos polegares.

As Tabelas 1, 2, 3 e 4 apresentam os resultados para cada categoria avaliada, descrita anteriormente. As porcentagens apresentadas nas Tabelas 1 e 2 indicam as taxas em que um dedo rotacionado ou não é corretamente classificado. As Tabelas 3 e 4 indicam as taxas em que casos para ângulos de rotação reais são classificados de acordo com a rotação predita. As diagonais representam os valores de verdadeiros positivos e são descritos nos gráficos de barras das Figuras 8(a) e (b).

No primeiro cenário, em que o classificador somente indica se ocorreu um mal-posicionamento devido à rotação ou não, a média das taxas de Verdadeiros Positivos foram de 90%, para todos os dedos, e 97%, para o polegar somente. Na detecção do ângulo de rotação (segundo cenário), a média das taxas de Verdadeiros Positivos foram de 90%, para todos os dedos, e 94%, para o polegar somente.

Vale observar que o polegar possui uma área útil maior, e também a projeção em diferentes vistas produzem claramente larguras diferentes, desta forma permitindo que os resultados da classificação do polegar sejam melhores. As taxas de classificação demonstram claramente esta característica para a análise do polegar.

5. CONCLUSÕES

A avaliação da qualidade tradicional para sistemas baseados em contato com o sensor, proposta pelo NIST, não pode ser totalmente estendida e aplicada para sistemas de impressão digital sem contato com sensor utilizando múltiplas vistas. Os sistemas sem contato tem características específicas, devendo ainda ser consideradas para se estabelecer uma avaliação da qualidade da impressão digital adequada. O fator mais importante, que se deve considerar, é em relação à completa ausência de uma superfície de contato, permitindo com que durante o processo de captura, o dedo do usuário possua maior quantidade de graus de liberdade.

Este trabalho apresentou uma metodologia baseada na utilização de redes neurais artificiais, que permite a detecção do mal-posicionamento dos dedos na captura de impressões digitais para sistemas multivistas sem contato. A idéia principal desta proposta concentra-se na definição de um ponto

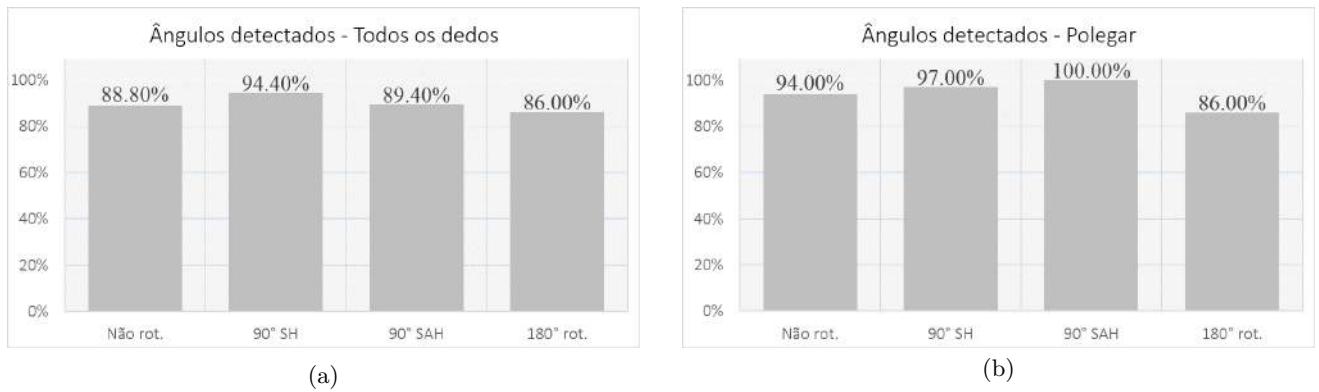


Figura 8: As porcentagens nas Tabelas 3 e 4 apresentam todos os resultados de acordo com os ângulos de rotação preditos. As diagonais que representam os verdadeiros positivos são apresentadas nos gráficos de barras em (a) e (b). A taxas médias de Verdadeiros Positivos em (a) e (b) são de 90% e 94%, respectivamente.

Tabela 3: Resultado da detecção de ângulo de rotação: todos os dedos.

		Predição			
Real		Não rot.	90° SH	90° SAH	180° rot.
	Não rot.	88.80%	0.50%	0.40%	10.30%
	90° SH	0.80%	94.40%	0.40%	4.40%
	90° SAH	1.60%	1.60%	89.40%	7.40%
	180° rot.	9.70%	3.40%	0.90%	86.00%

Tabela 4: Resultado da detecção de ângulo de rotação: polegar.

		Predição			
Real		Não rot.	90° SH	90° SAH	180° rot.
	Não rot.	94.00%	2.50%	0.50%	3.00%
	90° SH	0.00%	97.00%	3.00%	0.00%
	90° SAH	0.00%	0.00%	100.00%	0.00%
	180° rot.	3.50%	7.00%	3.50%	86.00%

de partida para que as metodologias de avaliação de qualidade para este tipo de sistema possa ser formalizada.

Os resultados obtidos pela metodologia proposta alcançaram os seus objetivos, permitindo que fosse detectado o mal-posicionamento dos dedos e polegares com 93% e 97.25% dos casos, respectivamente. Esta metodologia considera cenários em que o classificador pode verificar o ângulo de rotação de todos os dedos, com média de acerto de 90% para os dedos e 94% para os polegares.

Reforça-se também que a metodologia proposta não faz uso de qualquer pré-processamento nas amostras avaliadas e nenhuma das amostras coletadas foi descartada na montagem da base de dados. Desta feita, a base de dados foi construída utilizando um conjunto de imagens heterogêneas com as mais variadas características de iluminação, translação, borramento e contraste. Sendo assim, ficou demonstrado que a metodologia proposta é robusta e capaz de lidar com situações complexas em que estes fatores implicariam na qualidade da imagem capturada, proporcionando resultados

em que classificador utilizado fornece melhores resultados.

Em trabalhos futuros serão incluídos estudos de características complementares que permitam uma distinção mais refinada de detecção dos ângulos de rotação das imagens capturadas por este tipo de sensor. Além disso, outras métricas de qualidade podem ser consideradas e suas correlações com o mal-posicionamento rotacional dos dedos podem ser investigadas.

6. REFERÊNCIAS

- [1] R. Allen, P. Sankar, and S. Prabhakar. Fingerprint identification technology. In J. Wayman, A. Jain, D. Maltoni, and D. Maio, editors, *Biometric Systems: Technology, Design and Performance Evaluation*, chapter 2. Springer, London, 2005.
- [2] M. AlTarawneh, W. Woo, and S. Dlay. Objective fingerprint image quality assessment using gabor spectrum approach. In *Digital Signal Processing, 2007 15th International Conference on*, pages 248–251, July 2007.
- [3] H. Demuth, M. Beale, H. Demuth, and M. Beale. Neural network toolbox for use with matlab, 1993.
- [4] J. Dittmann and M. Hildebrandt. Context analysis of artificial sweat printed fingerprint forgeries: Assessment of properties for forgery detection. In *Biometrics and Forensics (IWBF), 2014 International Workshop on*, pages 1–6, March 2014.
- [5] M. Dusio, M. A. Olsena, and C. Busch. Fingerprint sample quality assessment via ridge line count using laplacian of gaussian edge finding. In *Biometrics and Forensics (IWBF), 2014 International Workshop on*, pages 1–6, March 2014.
- [6] J. Hammerle-Uhl, M. Pober, and A. Uhl. Systematic evaluation methodology for fingerprint-image quality assessment techniques. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, pages 1315–1319, May 2014.
- [7] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.
- [8] A. K. Jain and S. Pankanti. Automated fingerprint identification and imaging systems. In H. C. Lee and R. E. Gaensslen, editors, *Advances in Fingerprint Technology*, chapter 8. CRC Press, Boca Raton, 2nd edition, 2001.
- [9] R. Khanna and W. Shen. Automated fingerprint identification system (afis) benchmarking using the national institute of standards and technology (nist) special database 4. In *Security Technology, 1994. Proceedings*.

Institute of Electrical and Electronics Engineers 28th Annual 1994 International Carnahan Conference on, pages 188–194, Oct 1994.

- [10] A. Kumar and C. Kwong. Towards contactless, low-cost and accurate 3d fingerprint identification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3438–3443, June 2013.
- [11] R. Labati, V. Piuri, and F. Scotti. Neural-based quality measurement of fingerprint images in contactless biometric systems. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8, july 2010.
- [12] R. D. Labati, V. Piuri, and F. Scotti. A neural-based minutiae pair identification method for touch-less fingerprint images. In *Computational Intelligence in Biometrics and Identity Management (CIBIM), 2011 IEEE Workshop on*, pages 96–102, april 2011.
- [13] F. Liu, D. Zhang, C. Song, and G. Lu. Touchless multiview fingerprint acquisition and mosaicking. *Instrumentation and Measurement, IEEE Transactions on*, 62(9):2492–2502, Sept 2013.
- [14] D. Milgram. Computer methods for creating photomosaics. *IEEE Transactions on Computers*, C-24(11):1113–1119, Nov. 1975.
- [15] M. Olsen, E. Tabassi, A. Makarov, and C. Busch. Self-organizing maps for fingerprint image quality assessment. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 138–145, June 2013.
- [16] S. Pankanti, S. Prabhakar, and A. Jain. On the individuality of fingerprints. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(8):1010–1025, Aug 2002.
- [17] G. Parziale. Touchless fingerprinting technology. In N. K. Ratha and V. Govindaraju, editors, *Advances in Biometrics: Sensors, Algorithms and Systems*, chapter 2. Springer, London, 2008.
- [18] V. Piuri and F. Scotti. Fingerprint biometrics via low-cost sensors and webcams. In *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pages 1–6, 29 2008-oct. 1 2008.
- [19] J. Wayman. A definition of biometrics. In *National Biometric Test Center Colected Works 1997-2000*. San Jose State University, 2000.
- [20] B. Wing. Nist contributions to biometric technology. *IT Professional*, 16(2):38–44, Mar 2014.
- [21] J.-H. Xue and D. Titterington. t -tests, f -tests and otsu’s methods for image thresholding. *Image Processing, IEEE Transactions on*, 20(8):2392–2396, Aug 2011.
- [22] H. Yu and B. M. Wilamowski. Levenberg-marquardt training. In *Industrial Electronics Handbook, 2nd Edition*, volume 5, pages 12–1. CRC Press, 2011.

Um estudo comparativo das estratégias de fusão no nível de característica para Sistemas Biométricos Multimodais baseados em face e íris

Alternative Title: A comparative study of feature level fusion strategies for Multimodal Biometric Systems based on Face and Iris

Daniel M. M da Costa Universidade de São Paulo São Paulo - SP - Brasil daniel.moura.costa@usp.br	Henrique Passos Universidade de São Paulo São Paulo - SP - Brasil henrique.passos@usp.br
Sarajane Marques Peres Universidade de São Paulo São Paulo - SP - Brasil sarajane@usp.br	Clodoaldo A. M. de Lima Universidade de São Paulo São Paulo - SP - Brasil c.lima@usp.br

RESUMO

Com o avanço da tecnologia, novas abordagens para o reconhecimento automático da identidade de uma pessoa têm sido propostas e tal fato tem encorajado o emprego de Sistemas Biométricos. Essa abordagem utiliza características físicas ou comportamentais de uma pessoa para realizar a sua identificação. Os Sistemas Biométricos podem ser classificados como Unimodais ou Multimodais. Sistemas Biométricos Unimodais utilizam apenas uma modalidade biométrica para realizar o reconhecimento, ao passo que os Sistemas Biométricos Multimodais empregam duas ou mais modalidades. A construção de um Sistema Biométrico Multimodal pode ser realizada de diferentes formas, as quais são categorizadas de acordo com a sua arquitetura, nível de fusão e estratégia de fusão. O objetivo deste trabalho é investigar diferentes formas de fusão de modalidades biométricas no nível de característica, visando projetar um sistema multimodal com alto poder de reconhecimento. Neste artigo, nós utilizamos a Transformada Wavelet para extrair um conjunto de características de imagens de face e íris. Os resultados obtidos mostram que os Sistemas Biométricos Multimodais apresentam melhor desempenho que os Sistemas Biométricos Unimodais, em termos de taxa de reconhecimento calculada sobre a saída produzida pela Máquina de Vetores Suporte usada como classificador.

Palavras-Chave

Biometria Multimodal, Biometria Unimodal, Transformada Wavelet, Máquina de Vetores Suporte.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

ABSTRACT

With the technology advances, new approaches for automatic recognition of a person's identity have been proposed and such a fact has encouraged the use of Biometrics Systems. This approach uses physical or behavioural characteristics of the user in order to recognize or authenticate their identity. The Biometric Systems can be classified as Unimodal or Multimodal. The Unimodal Systems use a single biometric modality to perform the recognition, while the Multimodal ones use two or more modalities. A Multimodal Biometric System can be constructed in different ways, according to its architecture, fusion level and fusion strategies. The main of this work is to investigate and compare different feature level fusion strategies, in order to design a Multimodal Biometric System with high performance. In this paper, we used the discrete wavelet transform to extract the feature sets from iris and face images. Experimental results show that Multimodal Biometric Systems outperform Unimodal Biometric Systems according to recognition rate computed over the outputs produced by the induced Support Vector Machine classifier.

Categories and Subject Descriptors

I.2 [Artificial Intelligence]: Miscellaneous; D.2.8 [Pattern Recognition]: Application

General Terms

Design, Theory

Keywords

Multimodal Biometrics, Unimodal Biometrics, Wavelet Transform, Support Vector Machines.

1. INTRODUÇÃO

Devido à necessidade de fornecer segurança sobre a informação gerada em diferentes contextos, como por exemplo, numa transação monetária realizada pela internet ou

no controle de acesso a determinados locais, geralmente, são utilizadas estratégias como senhas, tokens, números PIN ou documentos pessoais para identificar o usuário e confirmar o seu acesso. Apesar dos métodos tradicionais obterem um bom desempenho na prática, estes estão sujeitos à fraude e à falsificação, visto que as informações podem ser roubadas, descobertas ou falsificadas por outros usuários [11].

Neste contexto, existe um crescente incentivo ao uso da tecnologia biométrica para melhorar, e até mesmo substituir, os métodos convencionais de segurança. A biometria emprega características físicas ou comportamentais do usuário para confirmar a sua identidade [10]. Uma das vantagens do uso dessa abordagem é que o usuário precisa interagir diretamente com o sistema, tornando a sua invasão mais difícil. As principais modalidades que podem ser usadas como biometria são a impressão digital [1], íris [24], face [15] e recentemente os sinais cerebrais [22] e cardíacos [16].

Para que uma determinada característica pessoal possa ser considerada uma modalidade biométrica, algumas propriedades devem ser satisfeitas [10] [9]: i) Universalidade: todas as pessoas devem possuir essa característica; ii) Distinção: por meio dessa característica deve ser possível distinguir duas pessoas quaisquer; iii) Mensurabilidade: deve permitir uma análise quantitativa da informação; iv) Desempenho: o reconhecimento deve ser rápido e com baixa taxa de erro; v) Não invasiva: não deve causar desconforto ao usuário durante a captura dessa característica.

Dependendo do contexto, um sistema biométrico pode atuar de duas formas diferentes [10], como um sistema de verificação (autenticação) realizando uma comparação um para um, cujo objetivo é confirmar ou negar a identidade reivindicada pelo indivíduo. Neste caso o sistema visa responder a seguinte questão: “Esta pessoa é quem ela diz ser?”. Já na identificação (reconhecimento), o sistema realiza uma comparação de um para muitos, o qual visa encontrar o cadastro pertencente ao indivíduo em um banco de dados. Neste o sistema é projetado para responder a questão: “Quem é esta pessoa?”. Para tornar possível o uso do sistema é necessário que o indivíduo primeiramente faça seu cadastro para que em um segundo momento possa ser realizado a requisição de autenticação ou identificação. Este artigo irá focar principalmente na identificação.

Um sistema biométrico é composto de quatro módulos principais [10], como visto na Figura 1: módulo sensorial, responsável por capturar os dados do usuário; módulo de extração de características, extrai a informação útil do dado capturado; módulo de comparação, compara a informação extraída com as armazenadas no banco de dados; módulo de decisão, recebe a saída do módulo anterior e decide quem é o usuário ou se o usuário será aceito ou rejeitado.

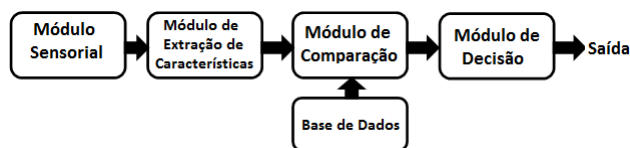


Figura 1: Arquitetura de um sistema biométrico.

Nos últimos anos, o reconhecimento biométrico tem alcançado melhorias em termos de confiabilidade e precisão, com algumas modalidades biométricas oferecendo um bom desempenho global. No entanto, mesmos os sistemas mais

avançados ainda enfrentam muitos problemas, tais como [19]: i) Dados de entrada ruidosos; ii) Variações intraclasse; iii) Similaridade interclasses; iv) Não-universalidade; v) Possibilidade de falsificação;

Visando atacar estes problemas, algumas abordagens têm sido proposta na literatura. Dentre estas abordagens destaca-se a Biometria Multimodal, que emprega diferentes modalidades biométricas em conjunto, como forma de gerar sistemas capazes de superar as dificuldades enfrentadas pelos Sistemas Unimodais. Biometria multimodal pode melhorar significativamente o desempenho de reconhecimento, além de aumentar a cobertura da população, impedindo ataques por repetição e reduzindo a taxa de falsa aceitação.

Para a construção de um Sistema Multimodal é realizada a união de dois, ou mais, sistemas biométricos unimodais. Essa união pode ser realizada de diferentes maneiras e são categorizadas de acordo com a sua arquitetura, nível de fusão e estratégia de fusão [10]. Em relação ao nível da fusão, as modalidades biométricas podem ser integradas no nível sensorial, extração de característica, comparação e decisão. De forma geral, a chave para o sucesso de um sistema biométrico multimodal está ligada a um esquema efetivo de fusão. Este trabalho irá investigar formas de integração no nível de característica de duas modalidades biométricas: face e íris.

Um dos principais módulos em um sistema biométrico é o módulo de extração de características. Neste módulo, após a etapa de pré-processamento das imagens de entrada, seja esta contendo a face ou a íris de um usuário, pode-se usar os pixels como entrada para um classificador. A vantagem de utilizarmos características extraídas das imagens, ao invés dos pixels, é que estes geram uma melhor separação interclasses e uma menor variação intraclasse representando melhor o domínio da aplicação. Dentre as diversas técnicas de extração de característica destaca-se a Transformada Wavelet (WT). Esta gera um conjunto de características, chamadas de coeficientes de Wavelet, que podem ser utilizados para representar a imagem original. Neste artigo, usaremos WT como extrator de característica.

De forma a projetar um sistema biométrico, unimodal ou multimodal, é necessário definir um classificador que será responsável por receber as características extraídas do usuário e verificar se o usuário tem, ou não, permissão para acessar um determinado recurso. Dentre os possíveis classificadores que podem ser utilizadas temos: Redes Neurais Artificiais [20], Máquinas de Vetores Suporte (SVM) [21] e K-Vizinhos Mais Próximos [12]. Aqui usaremos Máquinas de Vetores Suporte para compor o módulo de comparação do nosso sistema. A SVM [27] é um classificador binário que tem sido utilizado com sucesso em diversos problemas de classificação. De forma a garantir um bom desempenho ao usar esse classificador, alguns parâmetros devem ser escolhidos cuidadosamente, como a função *Kernel* a ser utilizada, seus parâmetros e o parâmetro de regularização.

O objetivo deste trabalho foi investigar diferentes formas de combinação de modalidades biométricas no nível de característica, visando projetar um sistema biométrico multimodal de alto desempenho. Duas modalidades biométricas foram utilizadas e como técnica de extração de característica foi empregada Transformada Wavelet, sendo adotada no nível de comparação às Máquinas de Vetores Suporte.

Esse trabalho está organizado nas seguintes seções: Seções 2 e 3 apresentam os conceitos e os métodos para construção de sistemas biométricos unimodais e multimodais, nas Se-

ções 4 e 5 são apresentadas uma introdução sobre Transformada Wavelet e Máquinas de Vetores Suporte, sendo essas utilizadas nos módulos de extração de características e de classificação, respectivamente. A Seção 6 descreve os procedimentos adotados para a realização dos experimentos, Seção 7 apresenta o desempenho obtido com os sistemas biométricos construídos. Finalmente, na Seção 8 é apresentado uma breve conclusão sobre os resultados e os próximos passos a serem adotados.

2. SISTEMAS BIOMÉTRICOS UNIMODAIS

Os Sistemas Biométricos Unimodais foram os primeiros a serem propostos na literatura. Estes empregam apenas uma modalidade biométrica para realizar o reconhecimento do usuário. Um Sistema Biométrico Unimodal baseado em face pode ser descrito como um sistema que usa uma câmera fotográfica para capturar uma imagem contendo a face do usuário e então verifica se o usuário está ou não cadastrado no sistema. Este é considerado um sistema de baixo custo de construção, já que qualquer câmera pode ser utilizada para capturar a face, porém quanto maior a resolução da imagem, mais informação pode ser extraída, facilitando o reconhecimento do usuário[6].

Apesar da imagem de capturada poder ser utilizada diretamente no sistema, geralmente, os sistemas empregam algoritmos de detecção de face para reduzir a interferência do fundo da imagem no desempenho do classificador. Diversos métodos têm sido propostos para detecção de face, dentre estes, podemos destacar o Algoritmo de Viola-Jones [29]. Este algoritmo emprega um conjunto de classificadores organizados em cascata para localizar a face em uma imagem. Neste trabalho foi utilizado esse algoritmo para destacar as faces e remover o fundo da imagem. Na Figura 2 temos um exemplo das faces detectadas e segmentadas.



Figura 2: Detecção de face usando o algoritmo de Viola-Jones.

Um dos grandes desafios enfrentados pelos Sistemas Biométricos baseados em Face está na detecção da face do usuário em ambientes com uma grande variação de iluminação, poses ou com a presença de acessórios [6]. Devido a essas limitações, os Sistemas Biométricos baseados em face são geralmente utilizados em ambientes controlados, que são locais onde esses algoritmos funcionam perfeitamente.

Já os Sistemas Biométricos baseados em íris empregam sensores específicos, ou câmeras de alta-resolução, para fazer a captura da íris do usuário. Apesar da íris não ser amplamente utilizada, essa é considerada uma modalidade biométrica de alto desempenho e confiabilidade, visto que o seu processo de geração é totalmente aleatório [23] e que a sua morfologia não sofre alteração, com exceção em casos específicos de algumas doenças [8].

Da mesma forma que os sistemas faciais, os sistemas envolvendo a íris, geralmente, empregam algoritmos de detecção e segmentação da íris de forma a separar a esclera e a pupila da imagem. Neste caso, é utilizada apenas a região da íris

para fazer o reconhecimento do usuário.

Para a detecção da íris utilizamos o Método de Wildes e Camus [3], que retorna um conjunto de pontos (x, y, r_p, r_i) , onde x e y são as coordenadas do centro da pupila, r_p é o raio da pupila e r_i o raio da íris. Utilizando esse conjunto de pontos e o Modelo de Daugman [5], podemos realizar a segmentação da íris em coordenadas polares. Sobre essa nova imagem é realizada o processo de extração de características. Um exemplo desse processo pode ser visto na Figura 3.



Figura 3: Processo de detecção e segmentação da íris.

3. SISTEMAS BIOMÉTRICOS MULTIMODAIS

De forma a superar as limitações enfrentadas pelos Sistemas Biométricos Unimodais, foram propostos os chamados Sistemas Biométricos Multimodais. Estes sistemas fazem o uso de duas, ou mais, fontes de informação biométrica.

De forma geral a área de Biometria Multimodal apresenta as seguintes vantagens em relação às abordagens unimodais: i) melhora significativa no desempenho de reconhecimento - diferentes modalidades medem informações complementares e por isto podem alcançar melhor desempenho; ii) melhora na robustez do sistema frente a variações do ambiente - todas as modalidades biométricas sofrem com algum tipo de interferência do ambiente, logo ao realizarmos a fusão de duas modalidades, espera-se uma redução no nível de interferência do ambiente; iii) diminui o problema da não-universalidade - ao utilizarmos mais de uma modalidade biométrica existe uma maior probabilidade de cobertura da população; iv) gera uma boa defesa contra ataques por convenção e repetição - o grau de dificuldade para que mais de uma modalidade biométrica seja falsificada ao mesmo tempo é maior, além disso essas características podem ser capturadas de forma aleatória.

Para a construção de um Sistema Biométrico Multimodal é necessário que algumas decisões sejam tomadas, como por exemplo, quais modalidades serão utilizadas, quais características serão extraídas de cada modalidade, em qual nível as características serão combinadas e qual será o método de fusão. Em [25] são apresentadas diferentes formas de construção de sistemas biométricos multimodais.

Após a extração das características de cada modalidade biométrica, é necessário que seja definido em qual momento será realizada a fusão e qual estratégia que será utilizada para fazer essa fusão. A Figura 4 apresenta as principais formas de fusão comumente empregadas. Em [25] essas estratégias são descritas detalhadamente. Neste artigo, iremos focar na fusão no nível de característica.

3.1 Estratégias para Combinação de Características

Em [7] foram propostas algumas estratégias de fusão, no nível de característica, de diferentes modalidades biométricas usando como técnica de extração de característica a

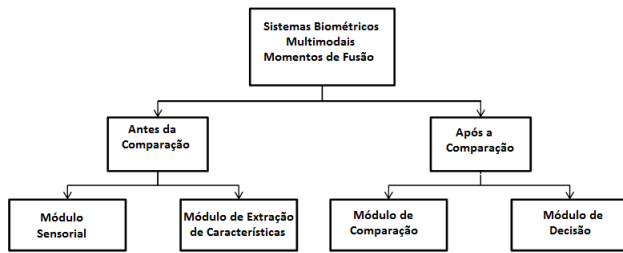


Figura 4: Níveis de fusão das características biométricas.

Transformada Wavelet. Basicamente, as estratégias propostas faziam a substituição de alguns dos coeficientes de Wavelet de uma determinada modalidade biométrica, usando alguma medida heurística, pelos coeficientes de Wavelet de outras modalidades. Com base nos coeficientes resultantes, uma imagem era construída e utilizada como entrada para o classificador. O problema desta estratégia é que a imagem reconstruída tinha a mesma dimensão da imagem original, resultando num conjunto de atributos de alta dimensionalidade. Visando contornar este problema, as estratégias de fusão adotadas neste artigo é realizada no nível de decomposição igual a 3, resultando em uma quantidade menor de coeficientes, sendo esses coeficientes utilizados como entrada para o classificador. As seguintes estratégias de fusão foram adotadas:

- *Estratégia #1* - nessa estratégia selecionamos os coeficientes de aproximação da face e concatenamos com os coeficientes de detalhes da íris;
- *Estratégia #2* - nessa estratégia selecionamos os coeficientes de aproximação da íris e concatenamos com os coeficientes de detalhes da face;
- *Estratégia #3* - nessa estratégia calculamos a média dos coeficientes de aproximação e concatenamos com máximo dos coeficientes de detalhes entre a face e a íris;
- *Estratégia #4* - nessa estratégia selecionamos o mínimo dos coeficientes de aproximação entre a face e a íris e concatenamos com máximo dos coeficientes de detalhes entre a face e a íris;

Após a aplicação das técnicas de fusão foram gerados novos vetores de características, sendo esses utilizados para representar a fusão da face e a íris do usuário.

4. TRANSFORMADA WAVELET

A Transformada de Fourier provê uma boa descrição das frequências de um sinal, mas não do seu comportamento temporal. Uma das primeiras abordagens para suprir essa carência da transformada de Fourier (e oferecer capacidade de análise no tempo e na frequência) foi subdividir um sinal em segmentos e calcular a transformada de Fourier em cada um deles. Essa abordagem é conhecida como Transformada de Fourier de Curto Tempo. A Transformada de Fourier de Curto Tempo utiliza uma função para segmentar o sinal original (criar janelas), e então aplica a transformada de Fourier em cada um dos segmentos gerados. Este método tem sido utilizado com sucesso em aplicações biomédicas.

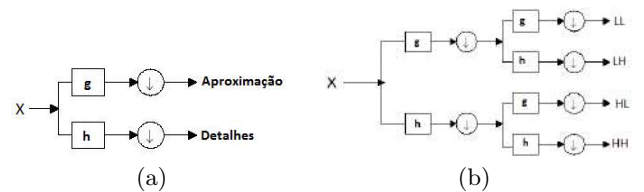


Figura 5: Processo de decomposição Wavelet (a) 1D e (b) 2D

Já a Transformada Wavelet descreve as propriedades de um sinal que muda durante o tempo [2]. Uma vantagem da Transformada Wavelet é que o tamanho da janela varia. Neste tipo de transformada, o sinal não é dividido em segmentos de tempo, mas em segmentos chamados de escala. Uma Transformada Wavelet permite analisar os dados em diferentes escalas, ou resoluções, as quais refletem as características mais ou menos detalhadas de um sinal. A análise de Wavelet consiste em adotar uma função-base e representar o sinal original como uma combinação linear de dilatações, contrações e translações (deslocamentos) sobre essa função-base. Existem várias famílias de funções-base Wavelets, dentre as quais destacam-se: Haar, Daubechies, Coiflets, Symlets e Morlet.

Ao invés de aplicarmos as *Funções Mães* para obtenção dos coeficientes de Wavelet, podemos aplicar um conjunto de pares de filtros FIR (resposta impulsiva finita) passa-alta H e passa-baixa G sobre o sinal. Após a aplicação destes filtros o sinal original é decomposto em duas partes, uma contendo os coeficientes de baixa frequência, representando os coeficientes de aproximação, e outra contendo os coeficientes de alta frequência, representando os coeficientes de detalhes.

De forma a utilizar a Transformada Wavelet em sinais 2D, tais como imagens, a ideia descrita acima pode ser expandida de forma a tratar as linhas e colunas da imagem como sinais independentes. Neste caso, o processo de decomposição é aplicado duas vezes seguidas, sendo que primeiro é realizado a análise nas linhas e a segunda nas colunas na imagem. Ao contrário da decomposição unidimensional, ao final desse processo geramos 4 sub-imagens diferentes contendo $\frac{M}{2} \times \frac{N}{2}$ coeficientes, onde M é o número de linhas e N de colunas da imagem original. Essas sub-imagens são chamadas de LL , LH , HL e HH , onde LL representa a aproximação da imagem original e as outras contêm os detalhes nas direções horizontal, vertical e diagonal, respectivamente. As figuras 5(a) e 5(b) apresentam o processo de decomposição 1D e 2D.

Após a Transformada Wavelet ser aplicada em uma imagem, ou sinal, é possível realizar o procedimento novamente sobre os sinais resultantes, gerando assim um novo nível de decomposição. Uma das formas de realizar esse processo é através do esquema de pirâmide [18]. Nesse artigo, aplicamos Transformada Wavelet para a extração das características com 3 níveis de decomposição, sendo que os níveis 2 e 3 foram gerados através da decomposição da sub-imagem LL do nível anterior. As sub-imagens resultantes do nível 3 (LL_3 , LH_3 , HL_3 e HH_3) foram concatenadas e utilizadas como vetor de características.

5. MÁQUINAS DE VETORES SUPORTE

A Máquina de Vetores Suporte (SVM) é conhecida como uma das principais técnicas a ser utilizada em problemas de classificação. SVM já foi utilizada com sucesso em aplicações envolvendo sistemas biométricos baseado em íris [13] e

face [14]. Essa técnica é indicada para problemas em que os dados de entrada apresentam alta dimensionalidade [17], uma vez que a sua formulação não depende diretamente da dimensionalidade dos dados.

Dado um determinado problema de classificação com um conjunto de treinamento $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, com entrada $\mathbf{x}_i \in \mathcal{R}^m$ e saída $y_i \in \{\pm 1\}$, considere um mapeamento não-Linear $\phi: \mathcal{R}^m \rightarrow \mathcal{R}^n$, onde n é muito maior que m , ou seja, os dados de entrada são mapeados para um espaço com maior dimensionalidade [26]. Nesse novo espaço (chamado de espaço de característica), a SVM busca construir um hiperplano de separação ótima, $(\mathbf{w}^T \phi(\mathbf{x}) + b)$ onde a distância entre duas classes é maximizada. Neste caso, $\mathbf{w}$ pode ser definido como a combinação linear de $\phi(\mathbf{x})$, isto é, $\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \phi(\mathbf{x}_i)$ [4].

Na formulação original das SVMs proposta por Vapnik [28], o hiperplano ótimo $\mathbf{w}$ pode ser encontrado minimizando a seguinte função:

$$\min_{\mathbf{w}, b, \xi_i} J(\mathbf{w}, b, \xi_i) = \frac{1}{2}(\mathbf{w}^T \mathbf{w}) + C \sum_{i=1}^N \xi_i, \quad (1)$$

sujeita às restrições:

$$y_i[\mathbf{w}^T \phi(\mathbf{x}_i) + b] \geq 1 - \xi_i, i = 1, \dots, N. \quad (2)$$

onde a variável C representa um termo de compromisso entre a complexidade do modelo e o erro de treinamento, e $\xi_i, i = 1, \dots, N$, são as variáveis de folga que medem a diferença entre a saída desejada e a saída da SVM. A solução para o problema de minimização da função-objetiva apresentada na equação 1, sujeito às restrições em 2, é dado pelo ponto de sela da função lagrangeana. O problema primal associado à minimização da função lagrangeana pode ser transformado em sua forma dual. O problema dual é dado por [28]:

$$\max_{\alpha} J(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \quad (3)$$

sujeito à $\sum_{i=1}^N \alpha_i y_i = 0$ e $0 \leq \alpha_i \leq C$, para $i = 1, \dots, N$, onde $\alpha_i (i = 1, \dots, N)$ são conhecidos como multiplicadores de Lagrange.

De forma a realizar o mapeamento $\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ na equação 3, não é necessário calcularmos $\phi(\mathbf{x}_i)$ ou $\phi(\mathbf{x}_j)$ de forma explícita, para isso podemos empregar as chamadas funções Kernel. Essas funções realizam o mapeamento não-Linear dos dados de entrada para um espaço de característica de maior dimensionalidade de forma eficiente [26], onde os dados são linearmente separáveis. Neste artigo utilizamos a função Kernel RBF, que pode ser descrita como:

$$K_{RBF}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right). \quad (4)$$

Após o processo de treinamento, as amostras que se encontram sobre a superfície de decisão apresentam valores de α_i , da equação 3, maiores que 0, sendo que essas amostras são chamadas de Vetores Suporte e são utilizadas para a definição da função de decisão do classificador [28]. Um determinado exemplo de teste $\mathbf{x} \in \mathcal{R}^m$ é classificado ao final de acordo com a saída de:

$$f(n) = \text{sign}[\mathbf{w}^T \phi(\mathbf{x}) + b] = \text{sign}\left(\sum_{i=1}^N \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b\right)$$

6. EXPERIMENTOS COMPUTACIONAIS

A seguir, nós fornecemos detalhes a respeito do conjunto de dados usado nos experimentos e como estes foram configurados. Em seguida, nós apresentamos a taxa de reconhecimento obtida para os sistemas unimodais baseado em íris ou face e para os sistemas multimodais usando fusão no nível de característica.

6.1 Base de Dados SDUMLA-HMT

Há poucas bases de dados públicas na literatura. Neste artigo, nós utilizamos a base de dados *SDUMLA-HMT Multimodal Database*. Essa base de dados é pública e contém as seguintes modalidades biométricas: Face, Íris, Veias do Dedo, Caminhar e Impressão Digital de 106 pessoas diferentes. No nosso experimento foram utilizadas apenas as 50 primeiras pessoas. O objetivo desta redução foi reduzir o tempo de processamento gasto na construção dos sistemas biométricos unimodais e multimodais.

A base de dados de face contém 84 imagens coloridas por usuário. As imagens foram capturadas com o usuário em 12 poses diferentes e cada pose contém 7 diferentes orientações. No total, a base de dados contém 8904 imagens. Na Figura 6 podemos verificar alguns exemplos de imagens de face dessa base. Nos experimentos realizados foram selecionadas apenas 10 imagens contendo o usuário em posição frontal com a câmera, de forma a simular um ambiente controlado em um Sistema Biométrico baseado em Face, onde grandes variações em relação à pose e orientação não são esperadas.



Figura 6: Exemplos de imagens de face da Base de Dados SDUMLA-HMT.

A base de dados de Íris contém 10 imagens por usuário, sendo 5 imagens representando cada íris (esquerda e direita), e capturadas a uma distância de 6 - 32 cm do dispositivo utilizado. A imagem é armazenada em escala de cinza com dimensão de 768×576 pixels. A base de dados contém 1060 imagens ao todo. Na Figura 7 podem ser visto alguns exemplos das imagens de Íris dessa base.



Figura 7: Exemplos de imagens de íris da Base de Dados SDUMLA-HMT.

Para a combinação de Face e Íris foi selecionada, de forma aleatória, uma das imagens de face e uma das imagens de íris, em seguida foi aplicada uma das estratégias de combinação

descritas na Seção 3.1.

6.2 Pré-Processamento das Imagens

De forma a preparar as imagens para o processo de extração de características, alguns passos foram realizados previamente para fazer a remoção de possíveis informações irrelevantes das imagens a serem utilizadas.

Inicialmente, as imagens de face foram convertidas para escala de cinza. Em seguida, aplicamos o Algoritmo de Viola-Jones de forma a realizar a detecção da região da face na imagem de entrada e reduzir a interferência que o plano de fundo pode causar ao sistema. Após a detectada a região da face, a imagem resultante foi ajustada para a resolução de 128×128 pixels. Posteriormente, a Transformada Wavelet foi aplicada para a extração das características usando 3 níveis de decomposição.

Em relação às imagens de Íris, também foi aplicada uma técnica de detecção e segmentação. Neste caso, nós empregamos o Método de Wildes e Camus para detecção e o Modelo de Daugman para segmentação da íris. Neste caso, a imagem polar foi gerada com a mesma dimensão da imagem de face, isto é, 128×128 pixels. Em seguida a Transformada Wavelet também foi aplicada sobre essas imagens usando 3 níveis de decomposição.

Após a aplicação da Transformada Wavelet usando 3 níveis de decomposição, independente da característica biométrica, utilizamos as imagens LL_3 , HL_3 , LH_3 e HH_3 para compor o vetor de características que será utilizado pelo classificador. Na Tabela 1 apresentamos as funções Wavelets utilizadas e a quantidade de características geradas por cada função Wavelet.

Tabela 1: Funções Wavelets utilizadas e a quantidade de características extraída.

Função Wavelet	Quantidade de Características
Daubechies 2	1296
Daubechies 4	1936
Symlet 3	1600
Symlet 4	1936
Symlet 5	2116

6.3 Configuração dos Experimentos

Para a construção dos Sistemas Biométricos utilizamos as SVMs para compor o Módulo de Comparação. As SVMs são um classificador binário e não podem ser utilizada diretamente em problemas com múltiplas classes. De forma a superar essa limitação utilizamos a abordagem um-contram. Neste caso, $\frac{nc*(nc-1)}{2}$ SVMs diferentes foram treinados representando todas as possíveis combinações de classes, nc representa o número de classes. A saída é definida utilizando voto majoritário.

Além disso para garantir que as SVMs possam ser analisadas em relação ao seu melhor cenário é necessário que os seus hiperparâmetros sejam sintonizados adequadamente. Neste caso, o parâmetro C que controla a complexidade do classificador e o parâmetro relacionado com a função de kernel utilizada, neste caso o Kernel RBF (equação 4). O parâmetro σ do Kernel RBF controla a abertura do mapeamento não-linear que é realizado sobre os dados de entrada. Baseado em testes preliminares, o valor do parâmetro C foi fixado em 1000, enquanto que o valor do parâmetro σ foi definido como 2^i , com $i = 0, 1, \dots, 9$.

Para melhor avaliar o desempenho do sistema biométrico construído foi utilizada validação cruzada 10-Fold. Nessa estratégia o conjunto de dados é separado em 10 partes diferentes, com a mesma quantidade de dados, onde uma parte é utilizada para teste e as 9 restantes são usadas no treinamento. Esse processo é repetido 10 vezes, a cada repetição uma parte diferente é escolhida como conjunto de teste. Este processo garante que todas as partes foram utilizadas para teste pelo menos uma única vez. O desempenho então foi avaliado de acordo com a taxa de acerto que o classificador obteve durante esse procedimento.

7. RESULTADOS

De forma a verificar a viabilidade e a capacidade dos Sistemas Biométricos Multimodais construídos usando as estratégias de fusão propostas na Seção 3.1, primeiramente realizamos a construção dos Sistemas Biométricos Unimodais usando Face ou Íris. Os resultados obtidos estão apresentados na Tabela 2. Para cada função Wavelet utilizada, apresentamos a taxa média de erro e o desvio padrão alcançado para o melhor cenário usando validação cruzada. Além disso, apresentamos o valor do parâmetro σ para o qual a SVM produziu melhor resultado.

Analisando os resultados alcançados utilizando somente os coeficientes de wavelet extraídos da Face, pode se perceber que, independente da função Wavelet utilizada, não foi possível obter uma taxa de erro menor que 10%. Neste caso, o melhor desempenho foi obtido usando a função Daubechies de ordem 2, que gerou uma taxa de erro de 16,20% com um valor de parâmetro do kernel (σ) igual a 128.

Com relação ao valor do parâmetro σ , podemos notar que, independente da função Wavelet utilizada, o melhor conjunto de valores de σ reside no intervalo de $[64, 128]$. Isto mostra que no caso da construção de um sistema biométrico envolvendo face, Transformada Wavelet e Kernel RBF, podemos escolher um valor dentro desse intervalo como um bom ponto de partida para o parâmetro do kernel. Comparando

Tabela 2: Resultados obtidos com os Sistemas Biométricos Unimodais usando Face e Íris.

Wavelet	Face		Íris	
	σ	Taxa de Erro	σ	Taxa de Erro
Daubechies 2	128	16,20%±6,69	256	7,80%±4,57
Daubechies 4	128	20,00%±5,81	64	10,00%±5,42
Symlet 3	128	17,40%±5,66	128	8,20%±3,46
Symlet 4	128	17,60%±5,94	64	9,80%±4,85
Symlet 5	64	19,20%±7,00	64	10,20%±6,43

as taxas de erro obtidas utilizando os coeficientes extraídos da íris com aquelas obtidas com relação à face, nota se que os sistemas baseados em íris alcançaram altas taxas de acerto. No melhor caso, foi obtido 7,80% de erro, com a função Daubechies de ordem 2. Para as funções Wavelet Symlet de ordem 3 e 4 a taxa de erro ficou abaixo de 10% de erro.

Já em relação ao valor do parâmetro σ podemos verificar que os melhores resultados para íris foram alcançados usando o parâmetro σ no intervalo $[64, 128]$, com exceção para o melhor cenário obtido que foi com o valor igual a 256 para função Daubechies de ordem 2. Apesar disso, podemos dizer que o intervalo $[64, 128]$ pode também ser utilizado como um ponto de partida para encontrar o melhor valor do parâmetro σ .

Os sistemas multimodais foram construídos utilizando as estratégias de fusão proposta na Seção 3.1, envolvendo a face

Tabela 3: Resultados obtidos com os Sistemas Biométricos Multimodais usando diferentes estratégias de fusão.

Função Wavelet	Estratégia 1		Estratégia 2		Estratégia 3		Estratégia 4	
	σ	Taxa de Erro	σ	Taxa de Erro	σ	Taxa de Erro	σ	Taxa de Erro
Daubechies 2	64	2,00%±1,63	128	1,60%±1,26	128	5,40%±2,31	64	1,60%±1,57
Daubechies 4	64	2,00%±2,66	128	3,20%±2,34	64	5,20%±3,15	128	2,60%±2,67
Symlet 3	64	2,60%±2,11	64	2,60%±2,83	128	4,20%±2,74	64	2,60%±2,11
Symlet 4	64	2,80%±2,69	64	3,00%±1,69	128	3,60%±2,63	64	1,60%±1,83
Symlet 5	64	2,40%±1,57	64	2,20%±1,47	128	4,00%±3,12	128	3,00%±2,35

e íris. Os resultados obtidos para cada estratégia são apresentados na Tabela 3. A primeira estratégia de fusão utiliza os coeficientes LL_3 extraídos da sub-imagem de face e os coeficientes LH_3, HL_3 e HH_3 extraídos da sub-imagem de íris concatenados para gerar o vetor de característica. Analisando a Tabela 3, pode-se notar que o desempenho obtido com essa abordagem conseguiu superar os melhores resultados obtidos pelos sistemas unimodais. Para esta estratégia, nós obtemos uma taxa média de erro menor que 3%, independente da função Wavelet utilizada.

O melhor resultado para a primeira estratégia foi alcançado usando a função Daubechies de ordem 2 e 4, na qual obtivemos uma taxa de erro de 2%. Para esta taxa de erro, o menor desvio padrão foi alcançado usando a função Daubechies de ordem 4. Usando a primeira estratégia, foi possível superar o desempenho alcançado pelos sistemas unimodais baseado em íris. Em relação à variação do parâmetro σ pode-se observar que em todos os experimentos para essa estratégia, os melhores resultados foram alcançados usando o valor igual a 64. Este valor ficou dentro do mesmo intervalo obtido para os sistemas biométricos unimodais.

A segunda estratégia utiliza os coeficientes LL_3 extraídos da sub-imagem da íris concatenada com os coeficientes das sub-imagens LH_3, HL_3 e HH_3 da face dos usuários. Novamente foi possível superar o desempenho alcançado pelos sistemas biométricos unimodais.

Usando a função Daubechies de ordem 2, a segunda estratégia produziu resultados superiores aos alcançados pela primeira e terceira estratégia. Sendo que, para o melhor cenário, foi possível obter uma taxa de erro média igual a 1,60%. Em relação ao parâmetro σ é possível verificar que o mesmo intervalo obtido pelos sistemas unimodais ([64, 128]) foi alcançado usando a segunda estratégia. Sendo que o melhor resultado foi obtido usando o valor de σ igual a 128.

A terceira estratégia utiliza a média dos coeficientes das sub-imagens LL_3 da íris e da face, e o máximo dos coeficientes das sub-imagens LH_3, HL_3 e HH_3 para compor o vetor de características. Semelhante às outras estratégias de fusão, foi possível superar o desempenho dos sistemas unimodais, independente da função Wavelet adotada.

Ao contrário das outras estratégias utilizadas e dos sistemas biométricos unimodais, essa estratégia obteve o melhor desempenho usando a função Wavelet Symlet de ordem 4. Neste caso, foi possível alcançar uma taxa de erro igual a 3,60% com o valor de σ igual a 128. Em relação ao valor do parâmetro σ , o mesmo intervalo adotado por todas as estratégias e os sistemas unimodais foi também obtido por esta estratégia, isto é [64, 128]. Por outro lado, o desempenho obtido com essa técnica foi o pior entre todas as estratégias de fusão adotada.

A última estratégia de fusão investigada aqui utiliza o mínimo dos coeficientes das sub-imagens LL_3 da íris e da face concatenados com o máximo dos coeficientes das sub-

imagens LH_3, HL_3 e HH_3 da íris e face. Essa estratégia forneceu o melhor resultado dentre todas as estratégias de fusão adotadas, com exceção da segunda estratégia. Entretanto, para algumas funções wavelets, o desempenho foi pior quando comparado as outras estratégias, conforme pode ser observado na Tabela 3. O melhor cenário foi obtido usando a função Daubechies de ordem 2, neste caso foi alcançado uma taxa de erro igual a 1,60%. Esse resultado também foi obtido usando a função Symlet de ordem 4, porém com um desvio padrão maior.

Analisando os valores do parâmetro σ , pode-se verificar que os melhores resultados em termos de taxa de erro de reconhecimento foram alcançados com valores de σ no intervalo [64, 128]. Esta mesma faixa de valores foi obtida para as outras estratégias de fusão e para os sistemas unimodais. Isto demonstra que esta faixa de valores contém bons candidatos para os valores de σ . Logo, ao utilizarmos a transformada Wavelet com 3 níveis de decomposição e imagens de entrada de dimensionalidade de 128×128 , um bom valor de σ pode ser escolhido neste intervalo.

De forma geral, podemos dizer que os sistemas biométricos multimodais, independente da estratégia de fusão, podem alcançar taxas de erro menores que aquelas obtidas pelos sistemas unimodais. No melhor cenário, os sistemas multimodais alcançaram taxa de erro igual a 1,60%, enquanto os sistemas unimodais alcançaram taxa de erro igual a 7,80%. Com relação a função Wavelet, independente do tipo de sistema (unimodal ou multimodal) e da estratégia de fusão, os melhores resultados foram alcançados com função Wavelet Daubechies de ordem 2. Por outro lado, a estratégia de fusão parece ser uma decisão importante na construção de sistemas multimodais. No caso das estratégias de fusão investigadas a taxa de reconhecimento incorreta variou de 1.60% a 2.0%. Já com relação ao parâmetro do kernel, apesar deste ser importante tanto para os sistemas unimodais quanto para os multimodais, para todas as estratégias o melhor resultado foi alcançado para valores de σ no intervalo [64, 128]. Acreditamos que este intervalo seja um bom ponto de partida para escolha do valor deste parâmetro.

8. CONCLUSÃO

Este trabalho realizou um estudo comparativo, em termos de taxa de reconhecimento incorreto, entre os sistemas biométricos unimodais e multimodais. As modalidades biométricas empregadas foram face e íris. Para a construção dos sistemas biométricos multimodais foi utilizada a fusão no nível de característica usando quatro estratégias diferentes. Para extração de característica foi utilizada a Transformada Wavelet.

Independente do tipo de estratégia de fusão, função Wavelet e modalidade biométrica, os sistemas biométricos multimodais alcançaram desempenho superior aos obtidos pelos sistemas unimodais. A estratégia de fusão parece ser um

ponto importante a ser levantado na construção de sistemas multimodais.

Como próximos passos, serão investigadas outras técnicas de extração de característica e qual o impacto dessas quando empregadas em um sistema multimodal. Além disso, serão investigadas outras formas de fusão, como por exemplo, fusão no nível sensorial e nível de decisão, e outros tipos de classificadores.

9. REFERÊNCIAS

- [1] A. Batool and A. Tariq. Computerized system for fingerprint identification for biometric security. In *Multitopic Conference (INMIC), 2011 IEEE 14th International*, pages 102–106. IEEE, 2011.
- [2] C. S. C. S. Burrus, R. A. Gopinath, and H. Guo. *Introduction to wavelets and wavelet transforms : a primer*. Upper Saddle River, N.J. Prentice Hall, 1998.
- [3] T. Camus and R. Wildes. Reliable and fast eye finding in close-up images. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 389–394 vol.1, 2002.
- [4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 1 edition, 2000.
- [5] J. Daugman. How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology*, 14:21–30, 2002.
- [6] K. Dharavath, F. Talukdar, and R. Laskar. Study on biometric authentication systems, challenges and future trends: A review. In *Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on*, pages 1–7, Dec 2013.
- [7] V. Divyaloshini and M. Saraswathi. Performance evaluation of image fusion techniques and its implementation in biometric recognition. *International Journal of Technology Enhancements and Emerging Engineering Research*, 2(3), 2014.
- [8] L. Flom and A. Saffir. Iris recognition system. U.S Patent 4 641 394, 1987.
- [9] L. Hong and A. Jain. Integrating faces and fingerprints for personal identification. *IEEE transactions on pattern analysis and machine intelligence*, 20:1295–1307, 1998.
- [10] A. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.
- [11] A. K. Jain, P. Flynn, and A. A. Ross. *Handbook of Biometrics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.
- [12] V. Kabeer, T. M. Thasleema, and N. K. Narayanan. Face recognition using state space parameters and k-nn classifier. In *Innovations in Information Technology, 2007. IIT '07. 4th International Conference on*, pages 476–480, Nov 2007.
- [13] H. F. Liau and D. Isa. Feature selection for support vector machine-based face-iris multimodal biometric system. *Expert Systems with Applications*, 38(9):11105 – 11111, 2011.
- [14] Z. Liu, L. Zhang, and L. Zhu. An improved face recognition method based on gabor wavelet transform and svm. In *Computational Intelligence and Design (ISCID), 2012 Fifth International Symposium on*, volume 1, pages 378–381, Oct 2012.
- [15] M. A. Lone, S. Zakariya, and R. Ali. Automatic face recognition system by combining four individual algorithms. In *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on*, pages 222–226. IEEE, 2011.
- [16] A. Lourenço, H. Silva, and A. Fred. Ecg-based biometrics: A real time classification approach. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.
- [17] A. Lourenço, H. Silva, and A. Fred. Ecg-based biometrics: A real time classification approach. In *Proceedings of the 22nd IEEE International Workshop on Machine Learning for Signal Processing*, 2012.
- [18] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(7):710–732, 1992.
- [19] V. M. Mane and D. V. Jadhav. Review of multimodal biometrics: Applications, challenges and research areas. *International Journal of Biometrics and Bioinformatics (IJBB)*, 2009.
- [20] M. Mazloom and S. Ayat. Combinational method for face recognition: Wavelet, pca and ann. In *Digital Image Computing: Techniques and Applications (DICTA), 2008*, pages 90–95, Dec 2008.
- [21] M. Mazloom, S. Kasaei, and H. A. Neissi. Construction and application of svm model and wavelet-pca for face recognition. In *Proceedings of the 2009 Second International Conference on Computer and Electrical Engineering - Volume 01, ICCEE '09*, pages 391–398, Washington, DC, USA, 2009. IEEE Computer Society.
- [22] R. Paranjape, J. Mahovsky, L. Benedicenti, and Z. Koles'. The electroencephalogram as a biometric. In *Electrical and Computer Engineering, 2001. Canadian Conference on*, volume 2, pages 1363–1366 vol.2, 2001.
- [23] H. Proenca and L. Alexandre. Iris recognition: An analysis of the aliasing problem in the iris normalization stage. In *Computational Intelligence and Security, 2006 International Conference on*, volume 2, pages 1771–1774, 2006.
- [24] D. M. Rankin, B. W. Scotney, P. J. Morrow, and B. K. Pierscionek. Iris recognition failure over time: The effects of texture. *Pattern Recogn.*, 45(1):145–150, Jan. 2012.
- [25] A. Ross and A. K. Jain. Multimodal Biometrics: an overview. pages 1221–1224, 2004.
- [26] B. Schölkopf and A. J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.
- [27] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
- [28] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
- [29] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Comp. Vision*, 57(2):137–154, 2004.

Influences of Organizational Culture in the Adoption of Agile Methodologies in Information Systems Development

A Systematic Mapping

Gerson Correia da Silva
Junior
Universidade Federal de
Goiás (UFG)
Goiânia, Brazil
gersoncorreia@outlook.com

João Paulo Amaral
Universidade Federal de
Goiás (UFG)
Goiânia, Brazil
joaopaulosi1@hotmail.com

Patrícia Gomes
Fernandes Matsubara
Universidade Federal do Mato
Grosso do Sul (UFMS)
Campo Grande, Brazil
patricia@facom.ufms.br

Valdemar Vicente
Graciano Neto
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
Universidade de São Paulo (USP)
São Carlos, Brazil
valdemarneto@inf.ufg.br

ABSTRACT

Information Systems (IS) industry has been pressured to deliver software products under reduced time-to-market and restrict budget. Agile Methodologies (AM) have been adopted to meet those aforementioned needs by the adoption of self-managing teams, iterative development cycles, fast delivery, and focus on functional software. Adoption of AM requires a suitable organizational environment, with significant changes in the employees' behavior and in the working processes, impacting, and suffering the impact of Organizational Culture (OC). In this direction, methods, actions, and policies must be thought to align the company's OC to the use of AM, in such a way that both can benefit each other. However, there is a lack of proposals that systematically investigate how to align both OC and AM. In this sense, a systematic mapping study (SMS) was conducted to identify studies that associate OC in initiatives of AM adoption. This paper presents the state-of-the-art about the influence of OC in the adoption of AM. The main contribution of this paper is bringing up an overview of the area, indicating perspectives of research, and exposing (i) a list of actions which have been reported to align OC and AM; (ii) OC factors recognized as essential to the successful adoption of AM; and (iii) the perceptions on a lack of awareness about the influence of OC in IT organizations.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

Categories and Subject Descriptors

H.0 [Information Systems]: General; K.4.3 [Organizational Impacts]: Management of Computing and Information Systems—*Project and People Management, Systems development*

General Terms

Information Systems, Systematic Mapping

Keywords

Organizational Culture, Agile Methodologies, Sociotechnical approach

1. INTRODUCTION

Information Systems (IS) software development is a complex activity. It inherently requires a sociotechnical approach. On one hand, social dimension comprises people involved in a project, the current culture in the organization, interactions among stakeholders, communication, and the entire spectrum of human aspects; while technical dimension includes a collection of processes, methods, techniques, and tools used to design the IS product.

Organizational Culture (OC) is a noteworthy social dimension. It corresponds to the set of assumptions and beliefs shared by a group [28]. It influences the work in the organization, impacting the routine, productivity, and delivery of results [23]. OC impacts employees routine, hierarchy, communication, teamwork, and it emerges when a set of assumptions are established by a group, becoming consolidated and recurrent in the daily work as the "correct form" of conducting the work.

On the other hand, time and budget pressures on Information Technology (IT) organizations have required the use of software development approaches which fastly deliver software [4]. In this context, Agile Methodologies (AM) have been adopted [2] for bringing a lighter approach to the

software development process, giving priority to social interaction, collaboration, and reduced documentation, meeting budget restrictions, time, and quality [1].

AM impose changes in the way people conduct their work. AM and OC are related and compose a bidirectional relationship where the cultural context can enhance some aspects of AM. On the other hand, AM can modify the cultural context in which it was inserted [18, 35]. Considering this mutual influence, it is necessary to propose approaches that align both perspectives of IS development, making OC and AM harmonically coexist in the organization. In this direction, a growing interest in the investigation of these areas has been perceived [5, 7, 14, 29]. Recently studies have been conducted to investigate how OC influences the adoption of AM, and how, conversely, these methods influence the culture [18, 22, 32, 35].

This paper presents results of a Systematic Mapping Study (SMS) that has been carried out to elucidate how OC and the adoption of AM influences each other in IS software development companies. For this, Section 2 presents some background; Section 3 briefly describes details on the methodology that guides the work, and reports the results; Section 4 brings some discussion; and finally, Section 5 presents conclusions and points out future work.

2. BACKGROUND

Agile Methodologies (AM) are distinguished from traditional software development methods by providing productivity, with focus on social interactions, adaptation to change, and constant delivery [9]. They provide strategies to organize teams and to prepare them for changes, enhancing communication with stakeholders [1], meeting quality and agility. In recent years, an increasing adoption of AM in the development of IS has happened due to the benefits it promises [2]. However, there is a low consensus of what can be considered agile and what is, in fact, agile. This concept depends on the context, and varies according to the perceptions of each group [26]. Indeed, the use of AM principles do not necessarily guarantee that the development will really exhibit such agility, since results are influenced by the way the team works. Thus, the use of AM can promote agility for some organizations, and not in others [18]. In this context, the concept of “emerging agility” [18] is considered an emergent property that may arise or not from the use of AM. It depends on the practices adopted in the organization. For the scope of this paper, the definition of ‘agility’ considered will be that one proposed by the Agile Manifesto [1], (i) focusing people and interactions over processes and tools; (ii) enhancing responsiveness to changes over following a plan; and (iii) enhancing collaboration with the client rather than the negotiation of contracts.

On the other hand, Organizational Culture (OC) is a major factor that also influences IS software development. OC is a framework of assumptions of a group, learned while solving problems along the years. This set of principles was enough well-succeeded to be considered valid, and to be taught to new members as the correct way of perceiving, thinking, and feeling in the organization [27]. OC is intrinsic to the relations within the company [16], influencing interpersonal relationships. It is usually stable and hard to change [27]. This can generate resistance when changes in internal processes are required [27].

Schein (1999) reports OC as a three-level stack. Figure

1 shows those levels: artifacts level, values level, and assumptions level [27]. The first level is the most exposed and easier to be identified. It consists of features that are visible/tangible, such as the arrangement of employees’ tables in the work space, their clothing, routines, and hierarchy. The second level is values. They materialize the way of acting shared by members of a group, such as ethics, working philosophies, rules of behavior, strategies, and objectives. Finally, the third level classifies individual and collective assumptions internalized by members of an organization, such as perceptions and feelings. The fundamental difference between second and third levels is related to perception: second level is more evident (as a behavior rule), while the third level is not usually evident, but influences on shared values (such as an apprehension feeling due to the pressure for results).

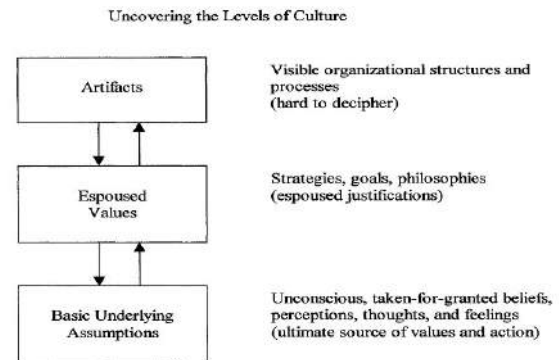


Figure 1: Levels of OC seconds Schein [28].

OC and AM can be considered quite orthogonal. However, they are related, and this relationship has an impact on the success of the adoption of AM in an organization [6]. Communication, for example, is an inherent OC factor, and a factor encouraged as practice in AM guides. Companies which adopt AM need to be flexible to changes, and flexibility is an OC factor. Then, there is a need for harmony between the methodology and the culture established in the organization. Thus, a convergence among OC and AM is relevant to study.

3. SYSTEMATIC MAPPING

A Systematic Mapping Study (SMS) was conducted to investigate the association between OC and AM. SMS is a literature review procedure to examine primary studies of a research area. From this analysis, data are collected to answer research questions [31]. SMS is guided by a protocol which specifies research questions, and inclusion and exclusion criteria followed to select studies and to validate the findings [20]. SMS brings many benefits, such as identifying future perspectives of research, as well as offering an extensive analysis of a research area. SMS is different from the systematic literature review (SLR) because SLR delivers its results on the analysis of very specific research questions. SMS, on the other hand, covers a research point under a broader perspective [21].

SMS was conducted to identify primary studies related to OC within the context of AM adoption. It was conducted from October 2014 to December 2014, and involved four researchers, one with expertise in OC, and all with relevant

knowledge in AM. We used principles proposed by Kitchenham and Peterson et al. [19] as reference to conduct this mapping. Briefly, the process was conducted following three phases: (i) planning, (ii) conduction, and (iii) reporting [8, 13].

3.1 Phase 1: Planning

During the planning phase, the protocol was defined. It guides the conduction of the SMS, and consists of the research questions, search strategy, selection criteria of primary studies, data extraction, and methods for results synthesis.

1) Research Questions To investigate how OC and AM influence each other, the following research questions (RQ) have been proposed:

RQ 1: Which set/collection of methods, techniques, actions, processes and/or activities have been used to align OC and AM adoption?

RQ 2: Which OC factors impact the success of the AM?

RQ 3: Organizations implementing AM have awareness about the influence of OC in their projects?

RQ 4: Which difficulties are reported by studies investigating OC in the context of AM?

2) Strategy Research The search string was prepared including the most common terms related to OC and AM, and their synonyms. The final search string is shown in Table 3.1. The following scientific bases were chosen to find the primary studies: ACM Digital Library¹, IEEEExplore², Science Direct³, Scopus⁴, and Web of Science⁵. Those bases were chosen due to the greater coverage and accuracy of results they offer, comprising a large number of conferences and specialized journals in IS area [11].

(agile AND (method OR methods OR methodology OR methodologies OR approach OR approaches OR practice OR practices OR principle OR principles OR process OR processes OR technique OR techniques) AND (culture OR cultural) AND (organizational OR organization OR organizations OR organisation OR organisations))

Table 1: Search String

3) Inclusion and Exclusion Criteria In order to select which studies are suitable for the scope of this mapping, the following inclusion criteria were determined:

IC 1: The study investigates the OC in the context of AM;
IC 2: The primary study presents detailed and well defined OC models;

IC 3: The study is based on previous works, which also addresses the OC in the AM context;

IC 4: Study presents research with results, validation, and justifications which reports the importance of OC regarding AM.

Exclusion criteria (EC) were prepared to eliminate studies that do not present contribution to answer the research questions. EC are presented below:

EC 1: Primary study does not even mention AM;

EC 2: Primary study does not even mention OC;

EC 3: Primary study was not written in english or portuguese;

EC 4: Primary study does not have full content, or it does not have an abstract;

EC 5: The primary study is directly related to another primary study of the same author. In this case, only the most recent primary study will be considered;

EC 6: The primary study consists of a collection of papers, for example, a conference or workshop;

EC 7: The primary study is not only available on a free-access base available in the place where the research was conducted.

4) Quality Questions In order to assess the quality of the selected studies according to the criteria previously established, the following quality questions (QQ) have been prepared based on a list of QQ commonly used in similar evidence-based studies [3, 12]:

QQ 1: There is a consistent reason for why the study has been conducted;

QQ 2: The authors present an overview of the related work and the area of the context in which the study is developed;

QQ 3: There is an adequate description of the context (industry, research environment, the methodologies used, and others) in which the study was conducted;

QQ 4: The study presents a clear justification for the methods used during its conduction;

QQ 5: There is a clear statement of contributions and data are presented to support them;

QQ 6: The authors explicitly discuss the credibility and limitations of their results;

QQ 7: The authors discuss the future work based on the contributions of the study.

Figure 2 shows the conduction process considering steps and products.

3.2 Step 2: Conducting

Primary studies were retrieved, selected, and evaluated according to the protocol. JabRef⁶, a free software for references management, was used to support the organization and analysis of primary studies. The selection of primary studies was divided into two steps: (i) Reading of the title and abstract, followed by a cross-validation of the results obtained by each reviewer, and when necessary, reading of the introduction and conclusion; and (ii) reading of the full study, and use of a form to perform data extraction. The evaluation of quality of included studies was performed and disagreements among researchers at any stage have been resolved in consensus meetings. The results of the search string applied in scientific basis are described in Table 3.2. *Science Direct* and *Scopus* were the basis with more returns, representing 21.45% and 34.76% of all returned studies, respectively. However, *ACM* and *Web of Science* were more precise, returning a higher percentage of the included studies.

3.3 Phase 3: Reporting

Our search resulted in 466 primary studies for analysis. After removing 137 repeated studies, 327 were submitted to selection step. After reading titles, abstracts, and, when necessary, the introduction and conclusion of each study, the results obtained were subjected to a cross-validation, and 14

¹<http://dl.acm.org/>

²<http://ieeexplore.ieee.org/Xplore/home.jsp>

³<http://www.sciencedirect.com/>

⁴<http://www.scopus.com/>

⁵<https://webofknowledge.com/>

⁶<http://jabref.sourceforge.net>

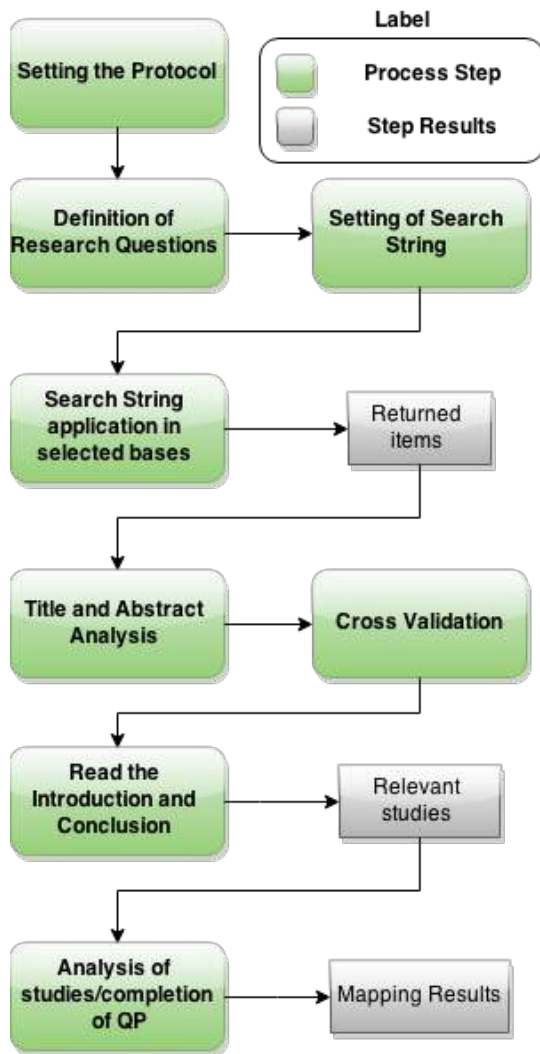


Figure 2: Systematic Mapping driving process of adapted Literature Tofan et al. [34].

primary studies were selected to be read in its entirety. Applying IC and EC, eight studies were chosen as included and reviewed. Table 3.3 presents the included studies and their respective years of publication. Table 3.3 shows selected studies. It brings original titles, as well as a brief summary of their contents.

RQ 1: Which methods, techniques, actions, processes and/or activities have been used to align OC and the use of AM?

Considering models to align OC and AM, only Kompella et al. (2014) [22] (from the eight included studies) explicitly present a set of recommended actions to align OC and AM. Their study presents an approach for changes that must be implemented in the organization. They consider people, processes, methodologies, and practices, and elaborate steps to follow in order to achieve results, and to benefit the emerging agility, aligning as result, the current OC and the AM adopted. They observed teams that use Scrum to reach their findings. Authors conclude that agility is influenced by the following list of factors related to OC: (i) strategies,

Basis	Returned	% Total	Selected
ACM	86	18,45%	4
IEEE	48	10,30%	1
Science Direct	100	21,45%	0
Scopus	162	34,76%	1
Web of Science	70	15,02%	2

Table 2: Statistics Used Primary Studies

ID	Study Included	Year
S1	Kerstin Siakas, Errikos Siakas [30]	2007
S2	D. Strode, S. Huff, A.Tretiakov [33]	2009
S3	Juhani Iivari, Netta Iivari [17]	2010
S4	Tolfo et al. [35]	2011
S5	Livari, Netta Livari [18]	2011
S6	Stavros Stavru et al. [32]	2012
S7	Juhani Dorairaj, Noble e G. Allan [10]	2013
S8	Kompella et al. [22]	2014

Table 3: Primary Studies List Included

plans and company's major goals; (ii) proposal of actions to achieve the planned goals; (iii) the growth in value delivery through the delivery of products; and (iv) training of employees to fulfill their actions.

In turn, Kompella et al. [22] prepare a process divided into stages in order to align the company's OC with the emerging agility. Their proposition is a training process to teach employees how to meet project's goals. There is not a validation step, since the conclusions come from observation. Furthermore, the study intention was providing *insights*, with no commitment of being a mature study. Their findings "suggest" that changes need to be performed to achieve the desired alignment, and a first step in this direction must be training the team. Additionally, they claim that an additional difficulty is choosing parameters to measure effectiveness of such an approach. They used customer satisfaction and productivity as parameters. However, they consider this a limitation since such parameters may vary according to the context where the project was conducted.

On the onther hand, Tolfo et al. [35] based their research on Schein's model and try to represent OC of three real companies. They mention the term "agile culture" as an ideal culture to adopt AM. Then, they characterize this agile culture using Schein's model divided into levels. Finally, authors mention that the OC of the analyzed companies are not compatible with the characteristics of the so-called agile culture. They conclude that aligning OC and AM consists of identifying the cultural levels of the organization and align them with the ideal agile culture. However, authors do not propose actions, activities, or models to achieve alignment between OC and AM. They only offer a first step in this direction.

Thus, we can deduce that there is still a necessity to consolidate a set of recommendations, procedures, processes or actions that can effectively achieve alignment between a company's OC and AM, reaching characteristics of the ideal agile culture. Proposals are still scarce and there is a significant lack of evaluation and validation.

RQ 2: Which OC factors impact the success of the AM adoption?

Regarding OC factors that impact the adoption of AM,

ID	Title	Summary
S1	The Agile Professional Culture: A Source of Agile Quality [30]	Comparisons between Schein, Mumford and agile principles in order to present another perspective to be considered for AM adoption.
S2	The Impact of Organizational Culture on Agile Method Use [33]	Case study of development projects that use some agile techniques, crossing results with the evaluation of a current OC. A set of OC factors to use AM are reported.
S3	Agile Methods and Organizational Culture: Reflections About Cultural Levels [35]	Three real companies are analyzed based on the Schein's model. Cultures and subcultures, and their influence on the adoption of AM are investigated.
S4	The Relationship Between Organizational Culture and the Deployment of Agile Methods [18]	A literature review on OC with AM.
S5	Organizational Culture and the Deployment of Agile Methods: The Competing Values Model View [17]	Presents theories and models discussed in other studies. addressing <i>Competing Values Framework</i> , and the <i>Theoretical Model</i> .
S6	Organizational Values and Agile Methods Deployment [32]	Proposes the creation of a model of OC that relates to AM from other models such as Schein.
S7	Agile Software Development with Distributed Teams: Senior Management Support [10]	Case study that investigates into the relationship between OC and AM, the challenge of managing a multinational multiple teams spread around the world working on the same project.
S8	Agile Methods, Organizational Culture and Agility: Some Insights [22]	An analysis of a company that adopts Scrum. The importance of teamwork is highlighted, focusing on communication, culminating in a process.

Table 4: Title and brief summary of each included study.

Strode et al. [33] identify such factors. From a list of nine factors, they survey the most common practices of AM, identifying six factors that contributes to link OC with the adoption of AM: (i) The organization fosters feedback and learning; (ii) Social interactions within the organization are reliable and collaborative, (iii) The project manager plays a facilitator role, (iv) The management style is of leadership and collaboration, (v) Teamwork in organization is flexible and participatory and encourages social interaction, (vi) the organization encourages training.

Siaskas and Siaskas [30] investigate the relationship between the types of culture proposed by Hofstede [15] (in a taxonomy which includes Democratic, Clan, Hierarchical and Disciplined), checking that one in which “agile culture” would fit more easily. According to them, AM should be considered a subculture within the existing OC, and under this perspective, the type of culture called democratic is characterized as the most appropriate, since it enhances factors such as flexibility and spontaneity within the organization’s hierarchy. Following this line of reasoning, we can conclude that one of the important factors that positively impact the adoption of AM is the type of culture existent. Those factors are similar to those found by Strode et al. [33].

Livari et al. [18] also identified specific types of culture that foster adoption of AM. They present hypothesis about a specific type of OC called “developmental” [25], that is more flexible and focused on adapting the organization to change. Finally, Livari and Livari [17] present results that summarize the main advances of some authors on the subject. Researchers claim that the relationship between OC and the use of AM need to be performed under the perspective of a specific model of OC (*Competing Values Model*). It has been initially proposed by Quinn and Rohrbaugh [25] to identify key indicators of effective business, and raise a number of hypothesis showing some barriers to the adoption of AM. Table 5 summarizes one of the remarkable taxono-

mies of OC and how they impact adoption of AM (positively or negatively). Data were collected from included studies. According to the table, cultures are classified as developmental, group, hierarchical, and rational. When analyzing if the adoption of AM is positively or negatively influenced by them, only hierarchical (one with a rigid hierarchy, and low levels of communication and flexibility) impacts negatively on AM adoption. Thus, the type of culture is also an important factor that influences and impacts the success of the AM adoption.

Culture Type vs Influence on AM	Positive	Negative
Developmental	X	
Group	X	
Hierarchical		X
rational	X	

Table 5: Types of Culture and the types of influence in AM proposed by Livari and Livari [18]

Table 6 compiles a set of factors that influence the OC in the adoption of AM, according to the studies analyzed. It was observed that, if the factors listed are present in the current OC, they receive and facilitate the adoption of AM, impacting positively. Similarly, if they do not exist in the current culture, this impacts negatively on the adoption of AM.

RQ 3: Organizations implementing AM have awareness about the influence of OC in their projects?

Two out of eight studies (S6 and S7) considered the assessed organizations unaware of the influence of OC in AM adoption. Remaining six studies, in turn, do not even mention if there is any awareness of the influence of OC on AM in the analyzed companies. In studies such as Tolfo et al. [35], in particular, authors report a large gap between what is

Studies	Organizational Culture Factors
S1, S5	Division of roles
S1, S3, S4	Management values innovation
S1, S2, S7, S8	Leadership encourages feedback
S2,S3 S4	Leadership takes risks
S2, S3	Organization values collaboration
S1, S2, S4	Organization allows flexibility of teams
S4, S8	Organization does not have a centralized hierarchy
S1	Satisfaction
S3, S8	Decision making is shared
S2, S3, S8	Teamwork is valued

Table 6: Factors that influence OC in adopting AM.

thought to be an agile culture and the practices really disseminated in the company. They analyze real organizations that follow agile values, and that references agile practices in its mission. They conclude that, when looking at deeper levels (second and third) of their OC, there is a misalignment with these values, and therefore an unawareness of the influence of such levels in the application of methodologies. Therefore, there is an indication that the alignment of the OC for AM adoption is still a little known practice. People are not awareness about the OC impact on AM adoption and in the companies routine as a whole.

RQ 4: Which difficulties are reported by studies investigating OC in the context of AM?

Volatility of term “agility” is one of the important difficulties reported by the studies. Two of them (S3 and S4) address more deeply what can be considered “agility”. Seven cite the Agile Manifesto’s conception of agility [1] as the main guide (only S6 does not mention Agile Manifesto). Three (S4, S5, and S8) deal with agility as an emergent property. From this perspective, agility arises from the interplay between the adopted methodology and working practices and culture already internalized in the organization. Thus, agility is a dynamic property that may emerge or not. In this direction, it is necessary to define a set of parameters that can be used to classify and evidence the emergence of “agility”, a difficulty reported by the studies. There is no consensus in defining what it is, and there is a tendency to consider as agile any flexible or adaptable development process.

In addition to the volatility of the term “agility”, other difficulties related to OC in projects that adopt AM are reported by studies such as (i) OC factors related to AM can not be generalized, since there are variables such as organizational environment and national culture that influence in each context; (ii) there is a complex and dynamic relationship between the OC and AM that poses additional challenges to organizations since continuous monitoring is necessary, identifying the influences from both parts of the OC in the AM, and AM in OC; (iii) studies using very strict techniques of scientific research (such as surveys for analysis of OC) may have a “gap” in its results, only identifying the first level of OC, which is usually contrasting with the deeper OC levels (second and third).

3.4 Threats to Validity

Table 7 brings the quality evaluation results of the studies

according to the criteria described in the protocol. Based on the results it is noteworthy that all the studies met the QQ1 and QQ2, and four (S1, S2, S4, and S8) of the studies did not meet the QQ7, i.e. they did not presented perspectives for future work, which does not impact reliability of the data collected. QQ6 identifies that there were studies that do not explicitly showed the limitations of their results. However, these data do not affect the reliability of the results.

Quality Issues	Yes	Some extent	No
QQ1	8	0	0
QQ2	8	0	0
QQ3	6	2	0
QQ4	7	1	0
QQ5	4	3	1
QQ6	4	3	1
QQ7	3	1	4

Table 7: Answers to Quality Questions.

We have also concerned with the validity of results of our systematic mapping. The following main threats to validity identified for this SMS were considered [24]:

- Missing of important primary studies: The publication databases used in this systematic study are considered the most relevant ones [12]. In addition, no limit was placed on the date of publications. However, some primary studies may have been missed;
- Reviewers reliability: None of the included studies were developed by research groups related to us. Therefore, we are not aware of any bias we may have introduced during the analysis. However, conclusions might have been influenced by reviewers’ opinions; and
- Data extraction: Since not all information was obvious to answer the research questions, some data had to be interpreted. Thus, to ensure the validity of this systematic mapping, discussions were conducted whenever a disagreement between the reviewers occurred.

4. DISCUSSION

Included studies present a variety of publication vehicles. Four were published in conferences (S2, S6, S7, S8), highlighting the International Workshop on Cooperative and Human Aspects in Software Engineering (CHASE), currently the most important discussion forum for OC in context of AM. One is a book chapter (S3), and the remaining three were published in journals (S1, S4, S5). In turn, publications origins are diverse as Brazil, Bulgaria, Canada, Greece, India, and New Zealand, each country contributing with only one study. Finland, distinctly, is represented by two studies.

Regarding the models to align OC and AM, only the study of Kompella [22] presented an explicit model, while Tolfo et al. [35] presented an indication of what could be a first step in this alignment. However, this model has not been validated effectively, and was applied only in the context of a specific AM (Scrum). Studies have established relations between these topics, without offering explicit actions for alignment and resolution of possible conflicts between the OC and AM adopted by organizations. Therefore, it is necessary to propose or adapt models, actions, and strategies that extend the results achieved, promoting more comprehensive approaches

that could be adopted by other AM, such as XP, Crystal, and Feature-Driven Development (FDD).

Considering OC factors that influences AM, despite the absence of specific actions, all evaluated studies raise at least one OC factor that has been confirmed to influence the adoption of AM, and this fact shows that factors such as lack of teamwork spirit, flexibility to change, and *feedback* are considered in most studies. Type of culture (considering those presented in OC taxonomies), and characteristics of the current culture are also important factors that impacts AM success.

Regarding the awareness of companies analyzed in studies on the influence of OC and AM, only two studies reported this kind of concern. Other studies are silent on this point. Therefore it can be assumed that there is need to disseminate awareness of OC and its importance in organizations. Additionally, the disparity between the intention to be agile and the reality in the company (the current OC) is highlighted.

Finally, the analysis of selected studies shows that there is still a lack of consensus on agility concept, especially taking into consideration emerging agility vision, what is considered a difficulty associated to the alignment of OC and AM. A massive reference to the Agile Manifesto [1] as a source of enlightenment to define AM was identified. Under this perspective, agility is an inseparable result of AM, a fallacious conclusion. By analyzing companies, studies conclude that adoption of AM does not guarantee that the organization reaches the desired agility. In most of the analyzed studies, the adoption of AM is given by the desire of the organizations to deal with frequently changing requirements, which often leads to short deadlines for decision-making and low quality results. The team that uses it must have characteristics that give support to the deployed process. Following this line of reasoning, Siaskas and Siaskas [30] identify that, considering the model of four types of OC proposed by Hofstede [15], the OC called democratic owns most of the characteristics aligned to the use of AM. As a corollary, the type of culture is an impacting factor on the success of AM adoption. Some of them have been considered more welcoming to AM approach, specially those with a high level of flexibility, intensive communication, team spirit, collaboration, and leadership management.

From the results we externalized the following research perspectives:

Concise definition of an ideal Agile Culture: IS development companies need to be investigated considering distinct types of AM, in order to characterize a concise, unified, and genuine conception of an ideal agile culture for AM adoption;

A unifying taxonomy for OC: Different names, OC types, and taxonomies are presented by the studies. An effort should be performed to unify those notations, creating a broad, complete, and coherent taxonomy for OC;

OC Alignment Models for AM: an investigation can be conducted to propose a set of models, practices, processes, techniques, and tools to support the alignment of OC with AM. Additionally, it is necessary to characterize the current OC in the organization and bring it closer to a culture that promotes agile values. In this sense, templates, checklists, and other procedures can be investigated to support this systematic alignment;

OC factors influencing AM: more studies are needed to

validate the results already found. It is necessary to investigate other factors that influence the OC in AM, as the National Culture. A mapping of specific factors of OC (as the influence of work space organization) that directly or indirectly influence in adopting AM is required, diagnosing the weight of each of the factors elicited in adopting AM, and how those factors influence the different types of AM adopted in companies such as Scrum, FDD, and TDD;

Awareness of OC in companies that adopt AM: studies can be conducted to investigate how companies applying AM have had access to OC concept, and the impact of this awareness in the company's daily routine. In this sense, research on the consciousness of the OC in AM can be conducted to understand the influence of OC. This can be used as an input to derive a set of decisions to modify the work environment. Those actions can support the achievement of characteristics to foster the adopted AM;

OC negative factors impacting on the adoption of AM: some characteristics of a current OC can be negative, such as inappropriate behavior, insubordination, immaturity of the teams, "culture of blame" [35], among others. A meticulous analysis can be performed to elicit factors that negatively influence the OC, directly impacting the success of AM adoption. The identification of these points can enable the adoption of contingency actions, motivating changes in the current OC.

5. FINAL CONSIDERATIONS

This paper presented results of a Systematic Mapping Study (SMS) which was carried out to investigate the influence of organizational culture (OC) in the context of IS software development companies which works with Agile methodologies (AM). Studies that investigate the relation between OC and AM were analyzed. As main contributions of this study, we can highlight the (i) compilation of a list of OC factors that are relevant and that effectively influence the success of AM; (ii) an overview of the OC area in the context of IS companies and studies dealing with AM; and (iii) research perspectives reporting unexplored gaps and research opportunities. These research directions involve the development of new models of action and/or practical methods, strategies, processes, guidelines, protocols, and other artifacts to assist IS software development companies to deploy AM without major difficulties, aligning the current OC to suitably fit agile practices. Such data and conclusions can support decision-making within a project or company. This study can contribute to those interested in this topic, serving as an starting point for future research and advances, in particular for the proposal of a model of actions to support a systematic proposal to align OC and AM. Additionally, this study contributes to the advancement of the state-of-art in research of a sociotechnical approach to enable IS software development in the Brazilian community, addressing a recurrent social aspect (OC), rising and correlating it with a disseminated technical practice of software development (AM).

6. REFERENCES

- [1] Agile manifesto. www.agilemanifesto.org, 2001. Acessado em 10 de janeiro de 2015.
- [2] Adoption of agile methods. <http://www.methodsandtools.com/dynpoll/>

- oldpoll.php?Agile2, 2008. Acessado em 5 de fevereiro de 2015.
- [3] M. S. Ali, M. A. Babar, L. Chen, and K.-J. Stol. A systematic review of comparative evidence of aspect-oriented programming. *IST*, 52(9):871–887, 2010.
- [4] Aoyama. M. agile software process and its experience. In *ICSE*, 1998.
- [5] K. Cameron and R.E. *Diagnosing and chaging organizational culture: Based on the competing values framework*. Reading MA: Addison-Wesley, 1999.
- [6] T. Chow and L. Cao. A survey of critical sucess factors in agile software projects. *JSS*, 2008.
- [7] A. Cockburn. *Agile software development*. Boston, Addison-Wesley., 2002.
- [8] G. Correia, J. P. Amaral, V. V. Graciano Neto, and P. G. F. Matsubara. Influence of organizational culture in the adoption of agile methods : Preliminary results of a systematic review (in portuguese). In *Proceedings of the 1st National Information Management Meeting*, ENGI '14, pages 1–6, Goiânia, Brazil, 2014. UFG.
- [9] E. G. da Costa Filho, R. Penteado, J. C. A. Silva, and R. T. V. Braga. Padrões e métodos ágeis: agilidade no processo de desenvolvimento de software. In *Proc. of 5th Sugarloaf Plop*, pages 156–169, 2005.
- [10] S. Dorairaj, J. Noble, and G. Allan. Agile software development with distributed teams: Senior management support. *ICGSE*, pages 197–205, 2013.
- [11] T. Dyba, B. Kitchenham, and M. Jorgensen. Evidence-based software engineering for practitioners. *Software, IEEE*, 22(1):58–65, Jan 2005.
- [12] T. Dybaa and T. Dingsoyr. Empirical studies of agile software development: A systematic review. *IST*, 50(9):833–859, 2008.
- [13] V. V. Graciano Neto, M. Guessi, L. B. R. Oliveira, F. Oquendo, and E. Y. Nakagawa. Investigating the model-driven development for systems-of-systems. In *Proceedings of the 2014 European Conference on Software Architecture Workshops*, ECSAW '14, pages 22:1–22:8, Vienna, Austria, 2014. ACM.
- [14] J. HighSmith. *Agile software development ecosystems*. England: Addison-Wesley, 2002.
- [15] G. Hofstede. Culture and organizations. *International Studies of Management & Organization*, pages 15–41, 1980.
- [16] C. I. *Administração nos Novos Tempos: Os novos horizontes da administração*. Campus: Rio de Janeiro, 1999.
- [17] J. Iivari and N. Iivari. *Organizational culture and the deployment of agile methods: The competing values model view*. Springer Berlin Heidelberg, 2010. cited By (since 1996)1.
- [18] J. Iivari and N. Iivari. The relationship between organizational culture and the deployment of agile methods. *IST*, 53(5):509–520, May 2011.
- [19] S. M. M. M. K. Petersen, R. Feldt. Systematic mapping studies in software engineering. *ICEAS*, 2008.
- [20] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. Systematic literature reviews in software engineering—a systematic literature review. *IST*, 51(1):7–15, 2009.
- [21] B. Kitchenham, R. Pretorius, D. Budgen, O. P. Brereton, M. Turner, M. Niazi, and S. Linkman. Systematic literature reviews in software engineering—a tertiary study. *IST*, 52(8):792–805, 2010.
- [22] L. Kompella. Agile methods, organizational culture and agility: Some insights. In *Proceedings of the 7th International Workshop on Cooperative and Human Aspects of Software Engineering*, CHASE 2014, pages 40–47, New York, NY, USA, 2014. ACM.
- [23] R. H. L. . S. M. M. H. Mohammad Jasim Uddin. Impact of organizational culture on employee performance and productivity: A case study of telecommunication sector in bangladesh. *International Journal of Business and Management*, 2013.
- [24] L. B. R. Oliveira, F. S. Osório, and E. Y. Nakagawa. An investigation into the development of service-oriented robotic systems. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 223–228, Coimbra, Portugal, 2013. ACM.
- [25] R. E. Quinn and J. Rohrbaugh. A spatial model of effectiveness criteria: Towards a competing values approach to organizational analysis. *Management science*, 29(3):363–377, 1983.
- [26] S. S. S. Sarker. Exploring agility in distributed information systems development teams: an interpretive study in an offshoring context. *ISR*, 2009.
- [27] E. H. Schein. *Organizational culture and leadership*. Jossey-Bass Inc, 1985.
- [28] E. H. Schein. *Organizational culture and leadership*, volume 2. John Wiley & Sons, 2010.
- [29] K. Schwaber and M. Beedle. *Agile software development with Scrum*. New Jersey: Prentice Hall, 2002.
- [30] K. V. Siakas and E. Siakas. The agile professional culture: A source of agile quality. *Softw. Process*, 12(6):597–610, Nov. 2007.
- [31] D. Šmite, C. Wohlin, T. Gorschek, and R. Feldt. Empirical evidence in global software engineering: a systematic review. *Empirical software engineering*, 15(1):91–118, 2010.
- [32] S. Stavru. Organizational values and agile methods deployment. In V. Cortellessa, H. Muccini, and O. Demirors, editors, *38th ECSEAA*, pages 282–285. IEEE Computer Soc, Los Alamitos, 2012.
- [33] D. Strode, S. Huff, and A. Tretiakov. The impact of organizational culture on agile method use. In *42nd HICSS'09*, pages 1–9, Jan. 2009.
- [34] D. Tofan, M. Galster, P. Avgeriou, and W. Schuitema. Past and future of software architectural decisions – a systematic mapping study. *IST*, 56(8):850 – 872, 2014.
- [35] C. Tolfo, R. S. Wazlawick, M. G. G. Ferreira, and F. A. Forcellini. Agile methods and organizational culture: Reflections about cultural levels. *J. Softw. Maint. Evol.*, 23(6):423–441, Oct. 2011.

Amadurecimento, Consolidação e Performance de SGBDs NoSQL - Estudo Comparativo

Alternative Title: Maturing, Consolidation and Performance of NoSQL Databases - Comparative Study

Vanessa Cristina Oliveira de Souza

Universidade Federal de Itajubá

Instituto de Matemática e Computação

Caixa Postal: 50 - CEP: 37500 903 - Itajubá – MG

55(35)3629-1827

vanessasouza@unifei.edu.br

Marcus Vinícius Carli dos Santos

Universidade Federal de Itajubá

Instituto de Matemática e Computação

Caixa Postal: 50 - CEP: 37500 903 - Itajubá – MG

55(35)3629-1827

marcusdcarli@gmail.com

RESUMO

Presencia-se ultimamente a expansão no número de dados gerados nas mais diversas áreas da computação. O modelo tradicional de banco de dados, chamado relacional não consegue lidar com este crescimento. O NoSQL surge como uma solução para esta problemática. Este trabalho teve por finalidade realizar um estudo sobre a arquitetura de alguns SGBDs NoSQL, evidenciar o motivo pelo qual eles conseguem lidar com grandes volumes de dados e mostrar algumas propriedades que o NoSQL baseia-se para fazer a gestão de dados. Para tanto, os SGBDs Redis e Cassandra foram utilizados e comparados ao banco relacional MySQL. Os resultados corroboram a literatura. O Redis apresentou melhor performance e o Cassandra a melhor escalabilidade. A maturidade e consolidação dos SGBDs NoSQL foram qualitativamente avaliados, de modo que este trabalho apresenta uma percepção de um usuário já familiarizado com a abordagem relacional dando seus primeiros passos com o NoSQL.

Palavras-Chave

NoSQL, escalabilidade, performance, consolidação

ABSTRACT

This paper presents a study on the NoSQL (Not Only SQL), a new dimension in databases. Lately witnesses an expansion in the number of generated data, and the relational model cannot deal with this data growth. The NoSQL comes as a solution to this problem. This paper aims to conduct a study on the architecture of DBMS, demonstrate why they can deal with this large volume of data and show some properties that NoSQL is based to make your data management. Therefore, DBMSs Redis and Cassandra were used and compared to the relational database MySQL. The maturity and consolidation of NoSQL DBMSs were qualitatively assessed on the parameters online documentation, software help,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26–29, 2015, Goiânia, Goiás, Brazil.

Copyright SBC 2015.

support virtual community and academic articles, so this work presents a perception of a user already familiar with relational approach taking their first steps with NoSQL. The results showed recognition of the relational database, but also showed a strong consolidation of NoSQL technology market and academic research environments, especially with Cassandra. The horizontal scale was tested with the DBMS Cassandra, who proved to be an excellent tool in this regard.

Categories and Subject Descriptors

H.2.4 [Database Systems]

General Terms

Performance, Design, Experimentation

Keywords

NoSQL, performance, horizontal scale, consolidation

1. INTRODUÇÃO

Com a expansão da Internet, o volume de dados cresce de forma exponencial e, nesse contexto, o *Big Data* vem ganhando grande importância. Em aplicações como redes sociais e sites de e-commerce, com inúmeros acessos e transações diárias, os sistemas de banco de dados relacionais tendem a ficar sobrecarregados. No modelo relacional[1, 2] os dados são estruturados, armazenados, manipulados e recuperados em forma de tabelas normalizadas, e todas as transações seguem propriedades de forte consistência chamadas ACID (Atomicidade, Consistência, Isolamento, Durabilidade).

Apesar da riqueza de recursos, os SGBDs relacionais tendem a aumentar a complexidade de utilização com o aumento do fluxo de dados. Eles não foram criados para trabalhar em ambientes distribuídos e o escalonamento, em geral, se dá pelo aumento da capacidade da máquina servidora (escalonamento vertical). Acontece que ao aumentar muito o fluxo de dados, o uso de *clusters* torna-se inevitável e o desempenho dos SGBDs relacionais cai.

Para suportar aplicações *Big Data*, onde é grande o fluxo de dados e a necessidade de escalabilidade dos servidores, surgiram os bancos de dados não relacionais, chamados NoSQL (*Not Only SQL*), que apresentam-se como uma otimizada solução onde os bancos relacionais têm grande deficiência.

De fato, o desenvolvimento da internet e da computação em nuvem fez surgir novos tipos de aplicações que destacou a tecnologia de banco de dados NoSQL. Aspectos como: escrita e leitura de alta concorrência com baixa latência, armazenamento e acesso eficiente de *Big Data*, alta escalabilidade e alta disponibilidade, baixo custo de administração e operação, apontam bancos NoSQL como uma forma eficiente e de baixo custo para suprir essa demanda.

[3] destacam a utilização de bancos NoSQL na computação em nuvem (*cloudcomputing*) e por empresas como IBM, Twiter, Facebook, Google e Yahoo!, com a finalidade de processamento analítico de *logs* Web, transações convencionais, entre outras tarefas. Bancos de dados NoSQL estão sendo adotados também em instituições financeiras, agências governamentais, comércio de produtos de varejo, entre outras áreas onde bancos de dados relacionais eram dominantes. Para se ter uma ideia de sua popularidade, o *ranking* dos sistemas gerenciadores de banco de dados mostra que bancos NoSQL já aparecerem entre os dez mais utilizados no mercado [4].

É importante ressaltar que o *Big Data* foi um estopim para a ascensão dos bancos NoSQL, mas como cita [5], ele não é o único motivo. A facilidade do uso de bancos NoSQL em ambientes distribuídos pode ser, entre outros, um motivo para a escolha desse tipo de *software*.

Dado esse contexto, o objetivo desse trabalho foi comparar três SGBDs, sendo dois NoSQL e um relacional. Como a tecnologia é recente, a intenção foi não apenas realizar testes de performance e escalabilidade, mas também verificar o amadurecimento e consolidação dos *softwares*, por meio de uma análise qualitativa de características como documentação, comunidades virtuais e artigos científicos. A hipótese testada é de que quanto mais completas e volumosas as características acima descritas, maior o amadurecimento e consolidação do *software* no mercado.

2. ESTADO DA ARTE

Segundo [6], *Big Data* descreve conjuntos de dados que crescem exponencialmente e que são demasiadamente volumosos, brutos ou desestruturados para serem analisados por meio de técnicas tradicionais de bancos de dados relacionais e de *business intelligence*. A *Big Data* expande-se através de três dimensões [7]:

- Volume – o tamanho dos dados é muito grande (na casa dos terabytes e petabytes).
- Velocidade – a velocidade que estes dados são criados é muito grande.
- Variedade – há uma grande variedade no formato dos dados, que inclui não apenas dados estruturados, mas também dados semi-estruturados e não estruturados de toda a espécie, como texto, áudio, vídeo, *posts*, arquivos de registros.

[8] apontam algumas limitações com relação à utilização dos bancos relacionais para armazenamento de *Big Data*.

- Leitura e escrita lenta - com o aumento do tamanho dos dados, é possível provocar bloqueios e outros problemas de concorrência, o que leva a um rápido declínio na eficiência da leitura e escrita.
- Capacidade limitada - bancos de dados relacionais atuais não suportam um volume grande de dados no motor de busca.
- Dificuldade de expansão - mecanismos de correlação entre múltiplas tabelas, devido ao uso de chaves estrangeiras

existentes em bancos de dados relacionais, não oferecem um bom mecanismo para a escalabilidade, sendo este um fator crucial para atender aos requisitos da *Big Data*.

De fato, [9] aponta a escalabilidade como a principal problemática dos bancos de dados relacionais para lidar com *Big Data*. Escalabilidade é a capacidade do sistema em lidar com o crescimento de dados uniformemente, podendo ser horizontal – que consiste no acréscimo do número de nós em um sistema – e vertical – relacionada à melhoria no hardware presente no sistema. Outra dificuldade inerente aos SGBDs relacionais para lidarem com *Big Data* é que os bancos relacionais são normalizados e seguem as propriedades de consistência ACID (Atomicidade, Consistência, Isolamento e Durabilidade). Somado a isso, há o fato de que, devido à grande quantidade de dados, indexá-los torna-se impraticável, pois o espaço ocupado pelos índices pode chegar a tamanhos críticos, o que leva a uma perda de desempenho.

Sendo assim, para atender a estas limitações, surgiram bancos de dados com arquiteturas e esquemas diferenciados, nomeados NoSQL (*NotOnlySQL*), capazes de processarem dados mais rapidamente, devido aos seus modelos de dados mais simples que os relacionais [8, 10].

Segundo [11], algumas características do NoSQL são: Não utilizam o modelo relacional e nem a linguagem SQL; são projetadas para rodar em *clusters*; tende a ser *open source*; e não possuem um esquema fixo, permitindo a persistência de em qualquer registro.

Os bancos NoSQL seguem a propriedade chamada BASE (*Basically Available, Soft-state, Eventual consistency*). Segundo [12], o BASE é o oposto de ACID: enquanto o ACID define que ao final de cada operação o banco deve manter sua consistência, o o BASE aceita que a consistência do banco de dados esteja em estado de fluxo, ou seja, tolera inconsistências temporárias para priorizar a disponibilidade.

A disponibilidade básica (*Basically Available*) determina que o sistema deve estar disponível na maior parte do tempo, mas que subsistemas podem estar temporariamente inoperantes. O estado leve e consistente (*Soft-State*) define que os desenvolvedores devem criar mecanismos para consistências de dados e a consistência eventual (*Eventual consistency*) garante que se certo registro não sofre atualizações, eventualmente estará consistente e todas as leituras retornarão à última atualização do valor. O período entre atualizações e o momento que ela é garantida é chamado de janela de inconsistência [12, 13]. A utilização do modelo BASE possibilita aos SGBDs NoSQL uma flexibilidade que atende as demandas dos ambientes distribuídos.

Uma característica importante de ressaltar dos bancos NoSQL é o teorema CAP [14], base da escalabilidade dos bancos de dados não relacionais. Acrônimo para *Consistency* (Consistência), *Availability* (Disponibilidade), *tolerance of network Partition* (tolerância ao particionamento de rede), o teorema descreve que uma aplicação distribuída não pode ter mais que duas propriedades simultaneamente, mas somente duas. Doze anos após a publicação do seu artigo sobre o teorema CAP, [15] percebeu que existem mudanças no cenário atual. Em sistemas onde o particionamento está presente e os projetistas devem fazer escolhas entre disponibilidade e consistência, umas das propriedades não precisa ser necessariamente negligenciada, mas sim apresentar uma combinação dessas. Sistemas NoSQL que fazem uso do particionamento tendem a focar primeiramente em

disponibilidade e deixar a consistência em segundo plano. No entanto, o sistema garante a consistência, mesmo que parcialmente. Hoje em dia, os projetistas devem escolher a que propriedades dar ênfase e não qual propriedade não usar.

Diferente do modelo relacional, os bancos de dados NoSQL possuem diversos modelos de dados que apresentam características que favorecem tipos diferentes de aplicações. Um banco NoSQL pode ser categorizado em quatro modelos diferentes, a saber: chave-valor; orientado a coluna; orientado a documento; e orientado a grafos. Detalhes desses modelos são amplamente encontrados na literatura técnica da área [8, 16, 17, 18, 19, 20, 21], entre outros.

Segundo [22], a utilização de bancos de dados NoSQL é uma alternativa a bancos de dados relacionais, pois provê consultas mais eficientes enquanto mantém flexibilidade e escalabilidade. Bancos de dados NoSQL tem a vantagem sobre a abordagem XML (*eXtensibleMarkupLanguage*) em relação a capacidade de fazer consultas, mas o XML tem a vantagem de reduzir o tempo de preparação para alimentar um banco de dados, principalmente quando é utilizado um banco XML nativo. O NoSQL também partilha desta característica, mas não é tão intuitivo quanto a abordagem XML.

Para [10], daqui para frente o foco do desenvolvimento NoSQL será em melhorar a compatibilidade com aplicações e com ferramentas de gerenciamento. Porém, a adoção do NoSQL em alguns casos será pequena, pois bancos de dados relacionais têm mais tempo de mercado e utilização, além de representarem grandes investimentos por parte dos vendedores e usuários.

3. METODOLOGIA

Este trabalho teve por finalidade realizar testes de performance em dois bancos de dados NoSQL e em um relacional, e revelar as impressões obtidas sobre a maturidade e consolidação dos *softwares*.

3.1 Material

Os SGBDs escolhidos neste trabalho foram o MySQL, Redis e o Cassandra. Utilizou-se também o *benchmark* YCSB. Para a definição dos SGBDs utilizados, levou-se em consideração o ranking dos SGBDs [4], que apontou o Redis e o Cassandra como os SGBDs chave-valor e orientado a coluna mais utilizados. A escolha do MySQL foi devido a sua popularidade e a presença de literaturas que o utilizam em diversas comparações com bancos NoSQL.

O Redis é um banco de dados chave-valor, *open source*, com a característica de que todas as suas operações são feitas em memória e periodicamente os dados são salvos em disco. Segundo [8], ao trabalhar na memória, o Redis garante uma boa performance e pode lidar com mais de cem mil operações de escrita e leitura por segundo. Porém, trabalhar em memória pode ser também uma desvantagem, por estar limitado pela memória do sistema. Segundo [23], o Redis não suporta indexação ou consultas complexas. Para operações básicas é muito rápido e simples, sendo o mais rápido dentre os bancos de dados chave-valor. Sua velocidade e estrutura de dados é uma combinação interessante que pode ser utilizada para a persistência de objetos simples. Estruturas de dados no Redis são chamadas de *Redis Datatypes*. Estas incluem *strings*, *lists*, *sets*, *sorted sets* e *hashes*. Segundo o modelo CAP, o Redis utiliza consistência e tolerância ao particionamento de rede.

Já o Cassandra é um banco de dados orientado a coluna, cujas características mais acentuadas são a escalabilidade e disponibilidade, sem a perda de performance. Foi criado pelo Facebook® e disponibilizado para comunidade em 2008. Ele possui esquema flexível, suporta consultas em range e possui alta escalabilidade [8]. Segundo [24], seu funcionamento consiste em um conjunto de um conjunto de colunas, chamado “supercoluna”, semelhante ao modelo chave-valor, no qual no campo valor pode existir outra chave que aponta para outra coluna. Atualizações são armazenadas em memória e posteriormente são escritas em disco, e as informações no disco são periodicamente compactadas. O Cassandra faz particionamento, replicação, detecção de falhas e recuperação automaticamente. No entanto, as cópias dos dados em outros nós são atualizadas assincronamente.

Segundo [25], o particionamento é uma característica chave do Cassandra. Um *cluster* com o Cassandra utiliza o mecanismo de *Gossip*, o que possibilita o controle de estados dos sistemas relacionados. A persistência de dados é feita em disco, todas as escritas em disco são sequenciais e um índice é gerado para buscas mais eficientes. Uma chave em uma família de colunas pode conter muitas colunas. Uma indexação especial é necessária para retornar colunas que estão muito longe da chave. Para evitar que se procure em todas as colunas, um índice de colunas é mantido no disco, o que permite um salto ao setor correto para reaver a coluna [25]. Em relação ao modelo CAP, o Cassandra utiliza disponibilidade e tolerância ao particionamento de rede.

Por sua vez, o MySQL é um banco de dados relacional que utiliza a linguagem SQL (*Structured Query Language*) e apresenta como pontos fortes a facilidade de manuseio, compatibilidade com muitas linguagens e suporte a praticamente todas as plataformas atuais.

Para os testes, fez-se uso do Cassandra versão 1. 12, Redis 2. 6. 7 e MySQL 5. 5. 32. Todos os sistemas foram testados utilizando a configuração padrão, instalados no sistema Linux Mint 15 64 bits. Para o Redis não foi colocado nenhuma restrição de utilização de memória RAM, ou seja, ele pode utilizar toda a memória disponível no sistema.

Segundo [26], *benchmarks* de um SGBD tendem a simular cargas de uma aplicação e gerar relatórios para uma comparação posterior. Esses *benchmarks* criam informações na aplicação que estão medindo o desempenho e simulam o acesso de usuários a esta aplicação. [27] definem o YCSB (*Yahoo! Cloud Serving Benchmark*) como um framework capaz de realizar *benchmarks* em diferentes sistemas gerenciadores de banco de dados, consistindo em um cliente gerador de cargas de trabalho com diferentes características, tais como cargas de trabalho com muitas leituras, muitas escritas e muitos escaneamentos. Uma das grandes vantagens do YCSB é a capacidade de criar ou adaptar cargas de trabalho.

O YCSB deve fazer várias escolhas aleatórias quando está gerando cargas, tais como: qual operação realizar (inserção, atualização, leitura ou escaneamento), qual registro ler e quantos registros escanear. Estas decisões ficam a cargo de distribuições. O YCSB tem várias distribuições, sumarizadas na tabela 1 [27].

Neste trabalho, utilizou-se o *benchmark* YCSB (sessão 3. 1. 1), com 50 threads para gerar os dados. Os testes foram realizados com as cinco diferentes cargas de trabalho (A, B, C, D e E).

Para o teste de escalabilidade do Cassandra foram utilizados quatro computadores (processador athlon1Ghz, 2 GB de RAM,

rede megabit) e a carga escolhida foi a C, por ser uma carga de somente leitura.

Tabela 1. Cargas de trabalho do YCSB-Fonte: [27]

Carga de trabalho	Operações	Seleção de registros	Exemplo de aplicação
A – Muitas Atualizações	Leitura: 50% Atualização: 50%	Zipfian	Ações recentes de gravação e armazenamento de uma sessão de usuário
B – Muitas Leituras	Leitura: 95% Atualização: 5%	Zipfian	Marcação de foto; adicionar uma tag é uma atualização, mas a maioria das operações é para ler as <i>tags</i> .
C - Somente Leituras	Leitura: 100%	Zipfian	Cache de perfil de usuário, onde os perfis são construídos em outros lugares.
D - Leitura utilizando Ultima	Leitura: 95% Inserção: 5%	Ultima	Atualizações de status do usuário, as pessoas querem ler suas últimas atualizações.
E – Pequenos Ranges	Escaneamento: 95% Inserção 5%	Zipfian/U niforme	Conversas por <i>threads</i> , onde cada varredura é para o <i>post</i> em uma determinada <i>thread</i> .

3.2 Métodos

Basicamente o trabalho se dividiu em três objetivos específicos diferentes: 1) Análise qualitativa do amadurecimento e consolidação dos softwares utilizados; 2) Testes de performance utilizando o YCSB e, 3) Teste de escalabilidade do software Cassandra.

A análise de amadurecimento e consolidação dos *softwares* avaliou qualitativamente os quesitos documentação *online*, *help* do *software*, suporte da comunidade virtual e artigos acadêmicos, comparando o Redis, Cassandra e MySQL.

Os testes de performance mediram o fluxo de dados e latências para 2000, 6000, 10000 e 14000 operações, utilizando o *benchmark* YCSB, com as cargas definidas na tabela 1. Os três SGBDs foram expostos ao teste.

O teste de escalabilidade foi realizado apenas com o SGBD Cassandra, pois o Redis utiliza replicação mestre/escravo, onde toda alteração no nó mestre é replicada aos nós escravos. O YCSB não atende bem esse cenário e por isso não foi testado. Já o MySQL, por ser relacional, não possui motivo para passar por um

teste de escalabilidade horizontal. O teste de escalabilidade utiliza o mesmo princípio do teste de performance, no qual se mede a latência de leitura para 2000, 6000, 10000 e 14000 operações, utilizando o *benchmark* YCSB, com as cargas de trabalho definidas na tabela 1. No entanto, para cada nó adicionado, realiza-se a carga e suas operações e, após as operações, outro nó é adicionado ao *cluster* e o processo repete-se sucessivamente para cada nó.

4. RESULTADOS E DISCUSSÃO

4.1 Amadurecimento e consolidação dos softwares testados

Com relação à documentação online, foram avaliados pontos como a facilidade para encontrar tais documentos, a clareza dos mesmos e o desenvolvimento desta documentação.

O Redis apresenta em seu próprio site uma documentação muito boa sobre o *software*, com tutoriais de instalação, lista de todos os comandos, tutoriais para configuração e links para livros tanto para iniciantes, quanto para usuários experientes. O Cassandra conta com uma documentação *online* que não é tão completa e objetiva. Não é hospedada em seu site, mas sim no site da *datastax*¹. A documentação do Cassandra vem sofrendo atualizações, inclusive algumas durante a escrita deste trabalho, que foram muito bem recebidas pela comunidade. O MySQL apresenta a melhor documentação *online* dentre os três bancos de dados estudados. Percebe-se que muitas pessoas contribuem com sua documentação. Os tópicos são mais bem organizados acarretando uma maior facilidade para encontrar a informação desejada. Essa documentação tão completa se deve ao tempo de mercado do *software* que é muito mais antigo que os outros dois apresentados neste trabalho.

Com relação ao sistema de ajuda do software (*help*), o Redis foi qualitativamente avaliado com o pior *help* entre os três SGBDs. Ainda faltam alguns elementos que os outros bancos de dados apresentam, como a apresentação de um exemplo demonstrando o comando requisitado e uma descrição geral deste. No geral é possível apenas sanar algumas dúvidas de sintaxe. O Cassandra apresenta um *help* considerado bom, melhor que sua documentação no site. Ele se aparenta muito com o *help* do MySQL, pois contém exemplos e explicações, ajudando muito os usuários inexperientes. O MySQL conta com o melhor *help* que, como já dito, possui explicações e exemplos, o que ajuda muito iniciantes e desenvolvedores.

No quesito de suporte da comunidade, foi avaliado o quanto as comunidades virtuais ajudam e o grau de comprometimento de seus participantes. O Redis possui a menor comunidade e talvez isso se deva ao fato dele ser chave-valor e ser simples. Mas é uma comunidade ativa, disposta a ajudar na solução dos problemas. Foi observado a falta de uma comunidade brasileira. No grupo de NoSQL do Facebook, não se encontra nenhuma discussão sobre o Redis. Pela importância do SGBD, espera-se que num futuro próximo surja uma comunidade nacional. O Cassandra, em conjunto com o MongoDB, são os SGBDs com a maior comunidade no ramo NoSQL. Como o Cassandra utiliza o modelo orientado a coluna, operações mais complexas podem ser realizadas com ele, o que gera mais dúvidas e discussões em

¹<http://www.datastax.com/documentation/gettingstarted/index.html>