Anais da IV Escola Regional de Sistemas de Informação do Rio de Janeiro

Proceedings of the IV Regional School on Information Systems of Rio de Janeiro

Rio de Janeiro, 25 e 26 de outubro de 2016
Rio de Janeiro/RJ - Brasil

Sociedade Brasileira de Computação (SBC)

Organizadores
Claudio Miceli de Farias (UFRJ)
Priscila Machado Vieira Lima (UFRJ)

Promoção
Universidade Federal do Rio de Janeiro (UFRJ)
Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais (NCE/UFRJ)

Realização
Sociedade Brasileira de Computação (SBC)

Editores

Claudio Miceli de Farias (Universidade Federal do Rio de Janeiro)
Priscila Machado Vieira Lima (Universidade Federal do Rio de Janeiro)

Título – Anais da IV Escola Regional de Sistemas de Informação do Rio de Janeiro

Local – Rio de Janeiro /RJ, de 25 e 26 de outubro de 2017

Ano de Publicação – 2017

Edição – 1ª

Editora – Sociedade Brasileira de Computação - SBC

Organizadores – **Claudio Miceli de Farias (Universidade Federal do Rio de Janeiro)**
Priscila Machado Vieira Lima (Universidade Federal do Rio de Janeiro)

ISBN: 978-85-7669-421-2

Apresentação

A sociedade vive um momento em que a tecnologia cada vez mais aumenta as possibilidades de se partilhar as funções cognitivas dos indivíduos através do suporte eletrônico. As organizações também são diretamente afetadas por esta nova realidade e, requerem a formação de profissionais que tenham condições de assumir um papel de agente transformador da sociedade, sendo capazes de induzir mudanças através da incorporação de novas tecnologias da informação na solução dos problemas.

É urgente a formação de profissionais com visão interdisciplinar, crítica, empreendedora, inovadora e humanística que possam viabilizar a busca por soluções computacionais para problemas complexos do dia-a-dia; considerando não somente questões técnicas relativas ao processamento da informação, mas também todo o contexto humanístico que abriga o problema em questão, é com essa perspectiva que se insere a proposta dessa escola.

Em 2017, a Sociedade Brasileira de Computação (SBC), a Universidade Federal do Rio de Janeiro (UFRJ) e o Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais (NCE/UFRJ) realizaram, entre 25 e 26 de outubro, no Espaço Flex do NCE/UFRJ a IV Escola Regional de Sistemas de Informação do Rio de Janeiro (ERSI-RJ 2017).

A IV ERSI-RJ teve como objetivo reunir estudantes, professores, pesquisadores e profissionais de Sistemas de Informação para promover discussões sobre temas relacionados a esta área. A programação da quarta ERSI-RJ contou com diversas atividades que se destacaram pelo seu alto nível técnico-científico e que cobriram diferentes aspectos, incluindo painéis com representantes da academia, indústria e governo; sessões técnicas para a apresentação dos artigos científicos selecionados e premiação dos melhores trabalhos, além de posters em diversas áreas.

Este ano tivemos cerca de 40 inscrições e 34 trabalhos foram submetidos (oriundos de diversos estados brasileiros), com 11 trabalhos completos e 8 posters aceitos para serem apresentados. Para que a IV ERSI-RJ tivesse este sucesso, foi necessário o trabalho de dezenas de pessoas, sejam coordenando as atividades, apresentando painéis e minicursos, participando dos comitês de

programas, na submissão de artigos ou na organização das atividades. A eles, o muitíssimo obrigado da organização geral da IV ERSI-RJ.

Agradecemos também à SBC e a UFRJ pela oportunidade e apoio. E, por último, mas não menos importante, agradecemos a todos componentes do grupo LABNET/UFRJ pela extraordinária dedicação ao evento e pela competência ímpar no auxílio na execução das atividades.

Nestes Anais, vocês encontrarão uma apresentação da escola, bem como dos trabalhos que foram apresentados. Aproveitem!

Em nome da Equipe Organizadora da IV ERSI-RJ 2017.

Claudio Miceli de Farias

Comitê Organizador do Evento

Coordenação Geral

Claudio Miceli de Farias (UFRJ)

Priscila Machado Vieira Lima (UFRJ)

Coordenação de Programa
Flávia Coimbra Delicato (UFRJ)

Sergio Manuel Serra da Cruz (UFRRJ)

Comitê de Programa

Adriana Vivacqua – UFRJ

Alessandro Copetti- UFF

Alexandre Correa- UNIRIO

Alexandre Sena – UERJ

Alvaro Robles Rincon – UFRJ

André Luiz de Castro Leal – UFRRJ

Angelica Dias – UFRJ

Claudia Cappelli – UNIRIO

Daniel Paiva – UFF

Daniel Schneider – UFRJ

Diego Brandão – CEFET/RJ

Ecivaldo Matos – UFBA

Eduardo Hargreaves – UFRJ

Emanuele Jorge – IFRJ

Fabiana Mendes – UnB

Fabricio Faria – UFRJ

Flavia Bernardini – UFF

Flavia Santoro – UNIRIO

Gabriel Aquino – UFRJ

Gizelle Vianna – UFRJ

Hugo Cesar Carneiro – UFRJ

Isabel Cafezeiro – UFF

Jonice Oliveira – UFRJ

José Ricardo Cereja – UNIRIO

Juliana França – UFRJ

Laci Mary Manhaes – UFF

Luci Pirmez – UFRJ

Luis Orleans – UFRRJ

Maria Luiza Campos – UFRJ

Mônica Silva – UFRJ

Rafael Costa – IFRJ

Rafael Escalfoni – CEFET-RJ

Raimundo Costa – UFRRJ

Robson Silva – UFRRJ

Rodrigo Monteiro – UFF

Tiago Cruz de França – UFRRJ

ARTIGOS TÉCNICOS

Trabalhos Premiados – Melhores Artigos

Pôster - Utilizando BPMN para Mapeamento dos Processos de Negócio de Instituições Públicas de Ensino
Leonardo Silva (UFS), Igor Nascimento dos Santos (UFS), Júlio Ribeiro (IFS), Fernando Oliveira (IFS) e Maria Carmo (IFS)

Trabalho completo - Mobile Recommendation System with Crowdsourcing and Geospatial Data
Kleyton Pontes Cotta (UFRJ), Raul Ferreira (UFRJ), Carlos Eduardo Barbosa (UFRJ) e Jano Souza (UFRJ)

Sessão de posters
Chair: Prof. Tiago França (UFRRJ)

Utilizando BPMN para Mapeamento dos Processos de Negócio de Instituições Públicas de Ensino - 1
Leonardo Silva (UFS), Igor Nascimento dos Santos (UFS), Júlio Ribeiro (IFS), Fernando Oliveira (IFS) e Maria Carmo (IFS)

Fiscaliza Cidadão: Uma Proposta para Promoção da Transparência e Participação em Políticas Públicas Através do Uso de Dados Abertos Governamentais - 3
Joathan Souza (FTA), José Silva (FTA), Gabriella Silva (FTA) e Everaldo Silva Neto (FTA)

An Userfriendly Approach for the Extraction, Transformation, Loading and Utilization of Heterogeneous Data in a Smart City Project - 5
Eliza Gomes (UFSC), Alexandre Ferreira (Polytech Grenoble), Anthony Geourjon (Polytech Grenoble), Jean-Francois Mehaut (Laboratoire dInformatique de Grenoble) e Mario Dantas (UFSC)

Proposta de Infraestrutura para Aplicações de SIoT para Universidade 7
Tiago França (UFRRJ), Ana Clara Silva (UFRRJ), Rodrigo Rodrigues (UFRRJ), Enaile Rebello (UFRRJ), José Orlando Gomes (UFRJ) e Jonice Oliveira (UFRJ)

Um sistema de recolhimento de dados para mapear informações edafoclimáticas em ambientes de fazendas verticais inteligentes 9
Emanuele Jorge (IFRJ), Jonice Oliveira (UFRJ), Claudio Miceli (UFRJ), Mário Sérgio Souza Pereira (IFRJ)

Utilizando a Computação Cognitiva para Classificar Denúncias da Linha Verde do Disque Denúncia do Estado do Rio de Janeiro 11
Ana Paula Camargo Pimentel (UFRJ), Alexandre Rangel (UFRJ), Eduardo Chiote (IBM), Walkir Brito (UFRJ) e Claudia Motta (UFRJ)

Explorando Computação Evolutiva em Workflows Científicos 71
Sergio Manuel Serra da Cruz (UFRRJ), Fabrício Firmino de Faria (UFRJ) e Anderson de Oliveira (CASNAV – Marinha do Brasil)

Sessão Técnica 3
Chair: Prof. Priscila Machado Vieira Lima

Uma investigação sobre estratégias a serem adotadas para o aprendizado de Inteligência Artificial no Ensino Fundamental por meio da Robótica Educacional 79
Rubens Queiroz (UFRJ), Fábio Ferrentini Sampaio (UFRJ), Priscila Lima (UFRJ)

Geração de Casos de Teste Independentes de Plataforma Utilizando Diagramas de Classes da UML Anotados com Restrições OCL 87
Marcos Furriel (UFRJ), Éber Schmitz (UFRJ), Mônica Silva (UFRJ) e Priscila Lima (UFRJ)

Sistema Computacional de Monitoramento de Qualidade de Água baseado em Arduino 95
Diego Brandão (CEFET/RJ), Raphael Guerra (UFF), Lucas Pinheiro (CEFET/RJ), Gabriel Gomes (UFRJ), Roberto Pontes (CEFET-RJ), Felipe Schubert Costa (CEFET- RJ), Gabriel Stefano (CEFET-RJ) e Henrique Junior (CEFET/RJ Unidade Nova Iguaçu)

Utilizando BPMN para Mapeamento dos Processos de Negócio de Instituições Públicas de Ensino

Leonardo de Jesus Silva[1], Igor Nascimento dos Santos[1], Fernando Lucas de Oliveira Farias[1], Maria do Carmo Bispo Silva [1], Júlio César Pacheco Ribeiro[1]

[1]Diretória de Tecnologia da Informação – Instituto Federal de Sergipe (IFS)
Campus Aracaju – Aracaju – SE – Brasil

{leonardo.silva,igor.santos,fernando.oliveira,carmo,julio.ribeiro}@ifs.edu.br

Abstract. *The processes carried out in public institutions sometimes suffer from the lack of standardization and transparency of the same, for internal and external community of the organization. In this sense, process modeling is performed using BPM concepts, being a managerial discipline integrating strategies and objectives of an organization and BPMN notation to solve such problems.*

Resumo. *Os processos realizados em instituições públicas, por vezes sofrem pela falta de padronização e transparência dos mesmos, para comunidade interna e externa da organização. Nesse sentido efetua-se a modelagem de processos utilizando os conceitos do BPM, sendo uma disciplina gerencial integradora de estratégias e objetivos de uma organização e a notação BPMN para solução de tais problemáticas.*

1. Introdução

A necessidade de alcançar metas e objetivos com efetividade se encontra presente nas organizações, na busca em atender seu público alvo da melhor forma [Brasil 2016]. A busca por modelos de gestão, que proporcionem uma gestão mais profícua dos recursos e que agreguem maior valor para o negócio bem como transparência dos processos [Dorneles et al.]. As instituições públicas brasileiras são prejudicadas pelo excesso de burocratização do setor público, o que por vezes gera problemas na excelência de serviços[Marchetti et al. 2009].

Constatando essa problemática, foi publicado um Guia de Gestão de Processos [1], o qual proporcionar uma integração das organizações públicas granular a burocratização existente. Esta proposta encontra-se no modelo ePING (Padrões de Interoperabilidade de Governo Eletrônico). Dentre as recomendadções, temos o Gerenciamento de Processos de Negócio (*Business Process Management* - BPM).

Este artigo propõe modelos de processo em instituições públicas com BPMN, realizando a exemplo, um estudo de caso em uma instituição federal de educação, a saber o Instituto Federal de Sergipe (IFS) e seus campus.

2. Metodologia de Modelagem de Processos

Para realizar a modelagem dos processos utilizou-se o Bizagi Modeler, o qual utiliza-se da notação BPMN, uma linguagem de diagramação para processos de negócio [Silver 2011].

[1]http://www.serpro.gov.br/menu/noticias/noticias-antigas/guia-de-gestao-de-processos-de-governo-e-lancado-em-maio

A Coordenadoria de Patrimônio (COPAT) identificou processos de negócio, estes efetuados pelas unidades gestoras, os quais deveriam ser modelados para evitar equívocos por parte das unidades gestoras em relação a execução destes; tendo como referência básica a instrução normativa, IN05/2013.

Neste trabalho foi utilizado uma combinação de técnicas de elicitação como: entrevistas, etnografia, JAD e pesquisas utilizando-se da técnica estabelecida por [StraussandJ and Corbin 1990], intitulada 5W2H *(Who?, When?, Where?, What?, Why?, How? e How Much?).*

3. Resultados e Discussão

Decorrente da metodologia apresentada na seção anterior, a modelagem do processo produziu o *workflow* a seguir. O processo de transferência interna que ocorre de um setor para outro, desde que estes estejam dentro do organograma da mesma unidade gestora. Este processo é inicializado, pelo setor ou unidade solicitante da requisição, e tem-se a participação do setor ou unidade destinatário da requisição da transferência. No entanto, o *workflow* de BPMN, apresentado na Figura 1, envolve mais um autor a Coordenadoria de Patrimônio (COPAT), além dos setores ou unidades envolvidos, para esta participar do processo e minimizar os possíveis riscos durante a execução do processo.

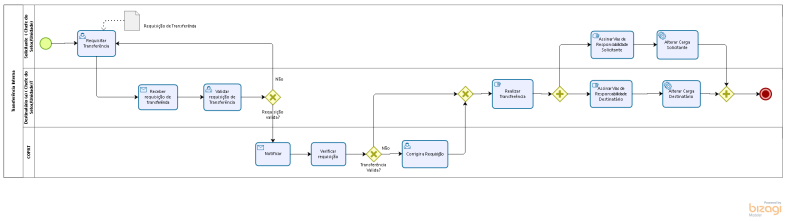


Figura 1. Workflow BPMN - Processo de Transferência Interna

4. Conclusão

Como resultado deste trabalho,iniciou-se uma uniformidade dos processos de negócio, os quais regem os modus operandi do IFS, estabelecendo modelos de processos para suas Unidades Gestoras e uma maior transparência para estes. Uma vez os processos mapeados, possibilita a otimização dos mesmos, em trabalhos futuros.

Referências

Brasil (2016). *Resolução Nº 45/2016/CS/IFS*. de 11 de abril de 2016, aprova ad referendum o Regimento Interno da Reitoria do IFS.

Dorneles, L., Deere, B. S. D. J., and FAHOR, C. P. Gestão organizacional: Elementos fundamentais para a organização de eventos.

Marchetti, C. d. C., Carvalho, R. d., and Mont'Alvão, C. (2009). A influência da gestão burocrática nas organizações públicas do brasil. *Revista INICIA*, 9(1):8–17.

Silver, B. (2011). Bpmn method and style, with bpmn implementer's guide: A structured approach for business process modeling and implementation using bpmn 2.0. *Cody-Cassidy Press, Aptos, CA*, 450.

StraussandJ, A. and Corbin, M. (1990). Basicsof qualitative research: Grounded theory proceduresandtechniques.

Fiscaliza Cidadão: Uma Proposta para Promoção da Transparência e Participação em Políticas Públicas Através do Uso de Dados Abertos Governamentais

Joathan Souza[1], **José Lucas da Silva**[1], **Gabriella Gomes**[1], **Everaldo Costa Silva Neto**[1]

[1]Faculdade de Tecnologia de Alagoas (FAT/AL) – Maceió – Alagoas

{joathanf, lucasnba2324, everaldocsneto}@gmail.com

Resumo. *Este artigo propõe o desenvolvimento do Fiscaliza Cidadão, uma aplicação Web que tem por objetivo utilizar dados abertos governamentais para fomentar a transparência e a participação do cidadão em políticas públicas, tornando-o fiscal na gestão do seu município.*

1. Introdução

No Brasil, a lei de acesso à informação, nº 12.527/2011, garante aos cidadãos o acesso à informação pública de forma aberta para que qualquer pessoa que tenha interesse possa fazer uso. Neste trabalho estamos interessados especificamente em dados de repasses federais aos municípios brasileiros. Os repasses federais são recursos públicos que são transferidos para cada município, eles são distribuídos por área (ex.: saúde, educação, assistência social). Em cada área existem os programas de governo (ex.: farmácia básica, unidades básicas de saúde), a qual um recurso é transferido.

Suponha que um cidadão (usuário) tenha o desejo de saber qual o valor transferido, em cada área, para seu município em um determinado mês, ou queira fazer uma análise mais detalhada observando quais programas, de uma determinada área, tiveram recursos transferidos, ou até mesmo queira fazer um comparativo de valores transferidos em um determinado período de tempo. Gerar esse tipo de informação pode ser um trabalho exaustivo para o usuário, uma vez que para ter essa informação ele irá precisar fazer várias buscas no Portal da Transparência[1] e depois fazer a integração dos resultados, ou poderá baixar o arquivo csv com os dados brutos, tratá-los e estruturá-los.

Considerando esse contexto e as dificuldades mencionadas acima, este artigo propõe o Fiscaliza Cidadão, uma aplicação *Web* que tem como objetivo coletar, estruturar, tratar e integrar dados de transferência de recursos federais para os municípios. O Fiscaliza Cidadão deverá permitir que os cidadãos (usuários) possam ter acesso, de forma mais intuitiva, a informações relevantes acerca dos recursos financeiros que chegam para seu município, de maneira que tais informações sirvam de subsídios para fiscalizar os seus gestores.

2. Fiscaliza Cidadão: Uma Proposta

O foco do Fiscaliza Cidadão é apresentar ao usuário, de forma rápida e fácil, informações analíticas acerca dos repasses federais recebidos em um determinado município. Para alcançar esse objetivo a aplicação está sendo desenvolvida seguindo o fluxo apresentado na Figura 1.

Os dados utilizados pela aplicação são **coletados** periodicamente, a cada mês, no portal da transparência. Em seguida, o processo de **ETL** descompacta o arquivo coletado,

[1]Local onde os dados citados estão disponibilizados (http://www.portaldatransparencia.gov.br/)

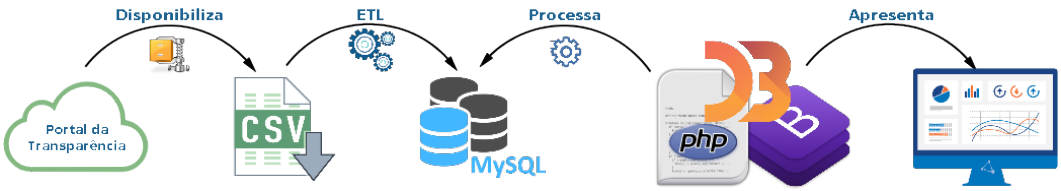


Figura 1. Fluxo do desenvolvimento da aplicação

estrutura e realiza o tratamento dos dados. Quando este processo é finalizado os dados são armazenados em um repositório gerenciado pelo SGBD MySQL. A **aplicação** está sendo desenvolvida utilizando a linguagem PHP, o *framework* Bootstrap, e a biblioteca D3.js para gerar as informações analíticas. Devido ao espaço resumido não há como fornecer mais detalhes da implementação.

Por fim, o usuário vai ter **acesso** a aplicação através de um *browser*, onde deverá selecionar o estado e município de interesse e poderá navegar através dos gráficos e informações apresentadas na tela, dentre elas destacamos: percentual de repasses por área, distribuição dos repasses por programas de governo, evolução (em valores/quantidade) de repasses em um determinado período de tempo. Além de relatórios mais específicos, como relação de favorecidos (em cada repasse - por área/por programa), entre outros.

3. Trabalhos Relacionados

O uso de dados abertos governamentais vem ganhando importância e gerando novas demandas, uma vez que a transparência das ações públicas é cada vez mais exigida pela sociedade. Nesse contexto algumas aplicações foram desenvolvidas, a qual destacamos a seguir: Meu Congresso Nacional [Brito et al. 2014] permite que cidadãos possam acompanhar o desenvolvimento dos parlamentares em sua função além de fiscalizar seus gastos; Painel dos Municípios [CGU 2017], oferece uma visão geral do município aos gestores e a sociedade através de informações relacionadas a fiscalização, transparência pública, ouvidoria, punições a empresas e demografia. A proposta apresentada neste artigo se diferencia das demais, pois é **específica para transferência de recursos federais** e está interessada em gerar informações analíticas, utilizando recursos estatísticos, através de gráficos e de relatórios específicos.

4. Considerações Finais e Trabalhos Futuros

Este artigo apresentou a proposta da aplicação Fiscaliza Cidadão, trata-se de um projeto de pesquisa que envolve o uso de dados abertos governamentais no contexto de *e-government*. O projeto já está em fase de desenvolvimento e tem previsão de conclusão em 12/2017. Como etapas futuras destacamos: (i) conclusão do desenvolvimento da aplicação; (ii) realização de testes para validar a aplicação; e (iii) realização de um experimento, com um conjunto de voluntários de diversos municípios, para avaliar o impacto e usabilidade da aplicação.

Referências

Brito, K., Santos Neto, M., Costa, M. A., Garcia, V. C., and de Meira, S. (2014). Using parliamentary brazilian open data to improve transparency and public participation in brazil. In *Proceedings of the 15th Annual International Conference on Digital Government Research*, dg.o '14, pages 171–177, New York, NY, USA. ACM.

CGU (2017). Painel dos municípios. <http://paineis.cgu.gov.br/index.htm> - acessado em agosto/2017.

An Userfriendly Approach for the Extraction, Transformation, Loading and Utilization of Heterogeneous Data in a Smart City Project

Alexandre Ferrera[1], **Anthony Geourjon**[1], **Eliza Gomes**[2], **Jean-François Mehaut**[1], **M.A.R. Dantas**[2]

[1]Polytech Grenoble
Saint-Martin-d'Hères, France

[2]Department of Informatics and Statistics (INE),
Federal University of Santa Catarina (UFSC),
Florianopolis, SC, Brazil

{alexandre78ferrera,anthony.geourjon}@gmail.com, eliza.gomes@posgrad.ufsc.br

jean-francois.mehaut@imag.fr, mario.dantas@ufsc.br

Abstract. *In this paper we present an implementation approach to a model of extraction, transformation and loading of heterogeneous data for smart city project. The main goal of our model is to provide to end user a friendly method of extracting, converting and sending data. Aiming to test the proposal it was utilized the ParticipACT Brazil project environment as a case study. This smart city project gets data from public and private companies, as well as crowd sensing campaigns, to propose some answers to ordinary urban challenges.*

1. Introduction

Big data can be understood as great quantity and variety of data. The capture and manipulation of these data can be a hard procedure due to the heterogeneity of formats and providers. To solve this, it is necessary to use data extraction and transformation tools [Gomes et al. 2016].

On this, we proposed the implementation of a model to extraction, transformation and loading heterogeneous data for smart cities projects. The main goal of our proposal is to provide a friendly interface capable of enabling that lay users can convert, send and receive data. We use the ParticipACT Brazil project [ParticipACT 2017] environment as our case study. This project proposes a smart city approach and, for this, it uses data from utilities companies and crowd sensing campaigns to direct manager to resolve urban problems, such as traffic jam.

Different data sources have different storage and privacy of their data. Therefore, we developed a software to be installed in the resources providers environment to simplify the extract, transform and load processes. Additionally, we developed a webservice with user-friendly interface to receive data from resources providers and storage them.

2. Proposal

We propose the implementation of an extraction, transformation and loading model. The main goal of our model is to facilitate conversion and sending of data from data sources, as well as, the receipt these data by users from different areas of knowledge.

This model is composed of two parts. The first part is on the data owner side, that is, each data sources have your ETL software version. The second one is on the ParticipACT Brazil environment, that is a webservice responsible for receive and storage data. As showed in Figure 1, our model is composed by 5 steps:

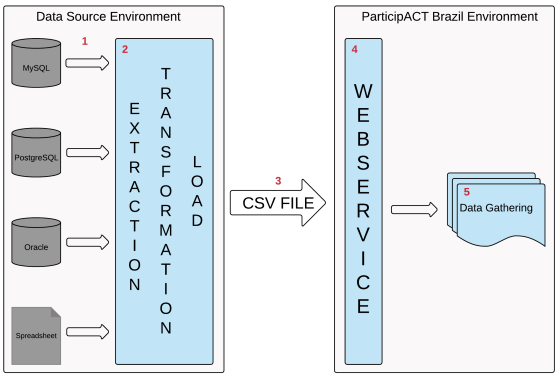


Figure 1. Proposed Model

1. The proposed model support several kinds of database, this way, the model can be applied to every company collaborating with the project.
2. The *Extract* step consists of accessing to the database and retrieving its content. To execute the *Transformation* step, is parsed the data extracted to transform it into the csv format. In the *Load* step, we gather the data so we can be prepared to send it in the next step.
3. The purpose of the *CSV File* step is to send the file loaded to the webservice inside of the ParticipACT Brazil Environment. This is handled by HTTP requests to the webservice.
4. The *Webservice* is divided in two main components: the API and the web interface. The first component receive and store data from HTTP requests in HDFS storage. The second is an user-friendly interface to monitor and manage the API.
5. The *Data Gathering* step consists of data storage location into ParticipACT Brazil environment.

3. Conclusions and Future Works

In this article we proposed the implementation of a model to convert and transfer, automatic and periodically, heterogeneous data for smart city project. As case study, we used the ParticipACT Brazil once this project receives data from different data sources (utilities companies and crowd sensing) with the objective of solving urban problems.

As future works we intend to insert security protocols for making transfers as well as implement and test our implementation in real case.

References

Gomes, E., Dantas, M., de Macedo, D. D., De Rolt, C., Brocardo, M. L., and Foschini, L. (2016). Towards an infrastructure to support big data for a smart city project. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2016 IEEE 25th International Conference on*, pages 107–112. IEEE.

ParticipACT (2017). Participact brasil. http://labges.esag.udesc.br/participact/. [Online]. Accessed in: Aug./2017.

Proposta de Infraestrutura para Aplicações de SIoT para Universidade

Rodrigo Rodrigues[1], Enaile Caldas Rebello[1], Ana Clara C. da Silva[1], José O. Gomes, Jonice Oliveira[2], Tiago C. de França[1]

Universidade Federal Rural do Rio de Janeiro Seropédica, RJ – Brasil

Universidade Federal do Rio de Janeiro Rio de Janeiro, RJ – Brasil.

rodsr98@gmail.com, jonice@dcc.ufrj.br, tcruzfranca@ufrrj.br

Resumo. *Este trabalho descreve uma proposta de infraestrutura para aplicações de SIoT para universidades. Tais aplicações podem ser criadas com o intuito de promover ambientes inteligentes por meio do uso de dispositivos eletrônicos e de pessoas. Esta proposta tem por motivação fornecer serviços comuns para aplicações de SIoT para campus universitários.*

1. Introdução

A Internet das Coisas Social (*Social Internet of Things* - SIoT) acrescenta à IoT uma abordagem social (de mídias sociais, por exemplo) [Atzori *et al.* 2012]. Nesse contexto e observando comunidades de universidades e mídias sociais (MS), Tabak *et al.* (2015) e Silva *et al.* (2016) propuseram uma aplicação para comunicação oportunística de SIoT (descrito pelos autores como "IoT *and people"*) com gamificação para incentivar a adesão e engajamento da comunidade universitária. Secron *et al.* (2016) usaram dados de MS, localização GPS e mapas para adicionar informações as rotas e informar sobre perigo relacionado a essas rotas. Alves et al. (2015) exploraram os conceitos de SIoT em campus focando apenas em aplicações rodando em dispositivos que se comunicam via NFC[1]. Este trabalho propõe uma infraestrutura para criação progressiva de soluções de SIoT para campus universitário. Aplicações de SIoT para universidades apresentam requisitos comuns e sua implementação para cada uma delas irá gerar retrabalho e dificultará a extração de informações usando os dados dessas aplicações. A criação progressiva diz respeito à inclusão (ou melhoria) de soluções sucessiva para campus universitário. Câmpus universitários (1) são ambientes comuns aos pesquisadores que reconhecem os problemas desse meio e (2) proporcionam um contexto de desenvolvimento de soluções que servem como modelos para cidades.

2. Descrição da Proposta

O campus universitário pode ser beneficiado por diferentes aplicações de SIoT. Por exemplo: indicação do tempo de espera para quem está em uma ponto de ônibus (uso GPS do ônibus e dos que estão nos pontos de paradas); monitoramento da temperatura, umidade e luminosidade para verificar as condições do ambientes de aula (ambiente quente, seco e escuro); deslocamento de pessoas no campus (identificação de picos de acesso ao restaurantes universitários, por exemplo); entre outros. Como principais características relacionadas a proposta é possível citar funcionalidades como: tratamento de fluxo contínuo de dados sendo publicados e consultados; a modelagem semântica dos dados (precisão do sensor, *timestamp*, tipo do sensor, aplicação de origem, usuário, publicação de usuários, etc.); API web para acesso aos recursos entre outros. A Figura 1 fornece uma visão geral da infraestrutura proposta. Nela estão os principais componentes para suportar diferentes aplicações de SIoT para campus

1 *Near Field Communication* (NFC) - https://www.iso.org/standard/56692.html

universitário. A infraestrutura observa a proposta de *gateway* de [Atzori *et al.* 2012]. Além da API, a figura mostra o componente Gerenciador de Dados que gerencia dados e ontologias necessárias. O Gerenciador de Regras contém a definição de como se relacionam os objetos enquanto o Gerenciador de Recursos gerencia a inclusão e manutenção de novos recursos que farão uso do *gateway*. A ontologia Lite-IoT[2] será a base para entendimento do relacionamento entre dados, dispositivos e serviços considerando a visão social a IoT. Este trabalho antecede os projetos citados como exemplo os quais estão em fase de concepção e servem como fomentador de requisitos para a presente proposta. A proposta é então a primeira fase de um projeto que pretende conduzir os câmpus universitários a campus inteligentes por meio do uso do paradigma da SIoT. Pretende-se testar, implantar e avaliar a proposta nos campus sede da UFRJ e UFRRJ. A parte social deste trabalho tem como principal ferramenta o CampusSocial [Tabak *et al.* 2015] que está em fase de implantação.

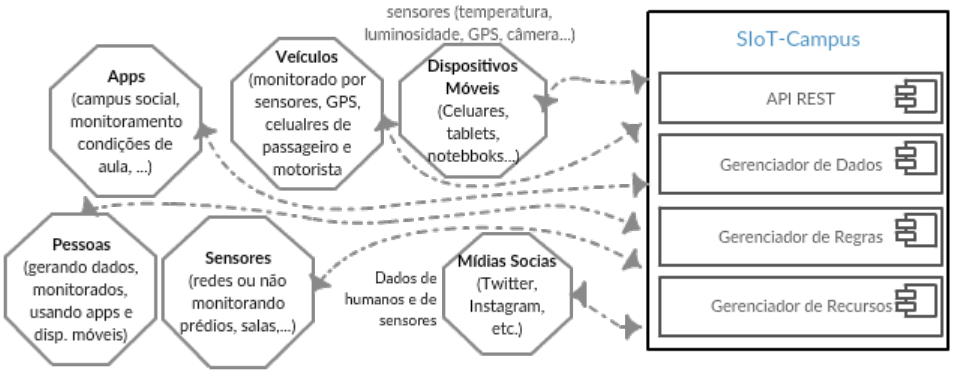


Figura 1: Principais componentes do gateway SIoT-Campus

3. Considerações Finais

Este trabalho é uma proposta de infraestrutura de suporte a criação de aplicações de SIoT para campus universitário. A infraestrutura proposta segue propostas de arquitetura para SioT e está sendo desenvolvida para tornar campus universitários inteligentes.

Referências

Alves, T. M., Andre da Costa, C., Da Rosa Righi, R. and Barbosa, J. L. V. (oct 2015). *Exploring the social Internet of Things concept in a univeristy campus using* NFC. In IEEE CLEI, Latin America.

Atzori, L., Iera, A., Morabito, G. and Nitti, M. (14 nov 2012). *The Social Internet of Things* (SIoT) – *When social networks meet the Internet of Things: Concept, architecture and network characterization*. *Computer Networks*, v. 56, n. 16, p. 3594–3608.

Secron, T. M., Roger, E. S., Farias, C. M. and França, T. C. (2016). SigaCiente: Uma ferramenta para inferência do trânsito e de rotas seguras baseada em dados sociais. In *III ERSI, Rio de Janeiro*.

Silva, E. R., França, T. C. and Oliveira, J. (2016). Aumento da Adesão e do Engajamento de Usuários do Campus Social com Uso de Mecanismos de Gamificação. In *III ERSI, Rio de Janeiro*.

Tabak, P., Figueiredo, E., França, T. C., Faria, F. and Oliveira, Jonice (2015). Campus Social: uma ferramenta para trocas oportunísticas de informações em campi universitários. In *Anais do CSBC*.

2 https://www.w3.org/Submission/iot-lite/

Um sistema de recolhimento de dados para mapear informações edafoclimáticas em ambientes de fazendas verticais inteligentes

Emanuele N.L.F. Jorge [1], Jonice de Oliveira Sampaio [2], Claudio Miceli de Farias[2], Mario Sergio de Souza Pereira[1]

[1]Instituto de Educação Ciência e Tecnologia – Instituto Federal do Rio de Janeiro – (IFRJ)
Avenida República do Paraguai, 120, Vila Sarapuí - Duque de Caxias
Rio de Janeiro – RJ – Brasil

[2]PPGI –Universidade Federal do Rio de Janeiro - NCE – (UFRJ)
Térreo, Bloco E, CCMN/NCE
Cidade Universitária -Caixa Postal 68.530 – cep 21941-590 – RJ – Brasil

emanuele.jorge@ifrj.edu.br, dcifrj22013.mario@proeja.com

jonice@gmail.com,claudiofarias@nce.ufrj.br

Abstract. *Population growth and increased rural exodus lead to concern about food insecurity. To solve this problem, solutions guided by the Internet paradigm of things, such as a smart vertical farm, are gaining more and more attention. In addition, we can use the Social IoT paradigm to consider the human being as a sensor of this network, considering it as a unit capable of generating data. This work does an initial research for the development of a data collection in an intelligent vertical farms and proposes a mechanism to recover the data and climatic information.*

1. Introdução

O crescimento demográfico e a urbanização estão intimamente associados ao desafio alimentar do século XXI, uma vez que não é possível aumentar a produção de alimentos em detrimento da expansão da área agrícola [Organization 2006]sendo portanto necessário o uso de áreas urbanas. As fazendas verticais [Buainain et al. 2016] [Despommier 2013] são ambientes urbanos transformados em plantações de pequenas culturas que possuem monitoramento contínuo [De Oliveira 2017] [Venkataraman 2008]. A tecnologia de monitoramento usada nesse contexto é a Internet das coisas (IoT) [Atzori et al. 2010] que pode ser considerada como uma rede mundial de objetos interconectados, de endereços exclusivos, com base em protocolos de comunicação padrão. De forma a considerar o conhecimento dos agricultores e sua relações sociais bem como as características das redes IoT pode-se tratar as fazendas Inteligentes como uma rede social de objetos inteligentes ou (SIoT - *Social IOT*), que conceitua as relações sociais entre objetos e seres humanos [Ortiz et al. 2014]. O objetivo deste trabalho é desenvolver um arcabouço de controle e decisão baseado em SIoT para o cenário agrícola.

2. Proposta

Esta proposta visa apresentar um arcabouço de controle e decisão que prevê a produção agrícola em uma fazenda vertical. Este arcabouço deverá ser capaz de lidar com grandes

volumes de dados e coordenar ações de diversas fazendas dentro da área urbana. A principal finalidade do arcabouço será prover soluções que auxiliem os agricultores através da análise de dados das fazendas inteligentes. Esses dados virão tanto do monitoramento vindo dos dispositivos IoT quanto das opiniões dos agricultores. Para isso, pretende-se aplicar técnicas de mineração de dados, análise de textos para identificar conteúdos postados e graus de interesse em um assunto/tópico [Ortiz et al. 2014].

Em um primeiro momento, propõe-se um mecanismo de monitoramento, controle e decisão das condições edafoclimáticas (clima e relevo) de uma fazenda vertical. Este mecanismo sugere ao agricultor as possíveis técnicas agrícolas a serem adotadas, as condições edafoclimáticas e os mecanismos de monitoramento e atuação implementados.

3. Considerações Finais

Este trabalho se propôs a apresentar o começo do desenvolvimento de um arcabouço de controle e decisão para fazendas inteligentes baseado em *social IoT* que permita a implementação de soluções de análise de dados. Para o desenvolvimento será necessário o levantamento das técnicas existentes para análise de dados na visão tradicional de IoT, o mapeamento das técnicas levantadas para o contexto de *social IoT*, o estudo das técnicas agrícolas a serem utilizadas em fazendas verticais bem como o planejamento e a executação de estudos experimentais visando avaliar a abordagem proposta.

Como trabalhos futuros buscar-se-á: (i) desenvolver a arquitetura do arcabouço; (ii) desenvolver novas técnicas de análise de dados; (iii) inferir com base no conhecimento dos agricultores e condições edafoclimáticas e (iv) contruir bases de dados de condições edafoclimáticas.

References

Atzori, L., Iera, A., and Morabito, G. (2010). The internet of things: A survey. *Computer networks*, 54(15):2787–2805.

Buainain, A. M., Garcia, J. R., and Vieira, P. A. (2016). O desafio alimentar no século xxi. *Estudos Sociedade e Agricultura*, 24(2).

De Oliveira, G. B. (2017). Uma discussão sobre o conceito de desenvolvimento. *Revista da FAE*, 5(2).

Despommier, D. (2013). Farming up the city: the rise of urban vertical farms. *Trends in biotechnology*, 31(7):388.

Organization, A. (2006). *The State of Food and Agriculture, 2006: Food Aid for Food Security?* Number 37. Food & Agriculture Org.

Ortiz, A. M., Hussein, D., Park, S., Han, S. N., and Crespi, N. (2014). The cluster between internet of things and social networks: Review and research challenges. *IEEE Internet of Things Journal*, 1(3):206–215.

Venkataraman, B. (2008). Country, the city version: farms in the sky gain new interest. *New York Times*, 15.

Utilizando a Computação Cognitiva para Classificar Denúncias da Linha Verde do Disque Denúncia do Estado do Rio de Janeiro

Ana Paula C. Pimentel[1], A. M. Rangel[1], Cláudia Motta[1], E. Chiote[1], Walkir Brito[1]

[1]Programa de Pós Graduação em Informática – Universidade Federal do Rio de Janeiro

pcamargo@unisys.com.br, amrangel@ufrj.br, eduardochiote@gmail.com, walkir.brito@gmail.com, claudiam@nce.ufrj.br

A pesquisa objetiva descrever uma solução para acelerar o processo de classificação das denúncias recebidas através da central de atendimento do "Disque Denúncia" do Estado do Rio de Janeiro. As denúncias são recebidas gerando um banco de dados não estruturados. Após o primeiro contato com a atendente, a denúncia é encaminhada por difusão para o setor responsável com classificação de prioridade de atendimento. A definição dessa prioridade é realizada de forma manual demandando tanto material humano quanto treinamento constante para apontar as dimensões de classificação e atribuições dos setores responsáveis. Para construir um modelo de classificação com IA foi utilizada a plataforma de computação cognitiva Watson IBM: API Classifier e um o método de classificação de riscos desenvolvido pela empresa Módulo Solutions for GRC [Módulo, 2013]. Para gerar uma rede semântica adequada para o sistema de classificação foi utilizado o banco de dados da Linha Verde do Disque Denúncia (2017) relacionado com o Código Penal Ambiental do Estado do Rio de Janeiro.

A plataforma Watson foi criada pela IBM para falar, ouvir, ver, interpretar e raciocinar como um cérebro humano. É uma máquina inteligente que interage na linguagem natural, reconhecendo, analisando padrões e aprendendo com as experiências passadas. Para que esse processo comece é necessário que regras generativas sejam implementadas com banco de informações de pertinência a função que a máquina irá exercer [DiMascio, 2016].

A API Classifier tem como objetivo central classificar dados não estruturados conforme o modelo construído pelo usuário. Ela utiliza algoritmos de aprendizado de máquina (*machine learning algorithms*) e redes neurais complexas para classificar entradas de textos curtos. É um classificador de linguagem natural, que aprende/treina com dados (textos, sequência de palavras) de exemplo e é capaz de classificar dados "não treinados" com base nesse aprendizado [IBM BLUEMIX Docs, 2017].

O processo é estruturado da seguinte forma: 1. Preparar os dados para treinamento com uma rede semântica pertinente (identificar o rótulo da classe; coletar textos e palavras representativas; combinar as classes com os textos); 2. Criar e treinar o classificador (usar a API para subir os dados para treinamento e iniciar treinamento); 3. Criar query para o classificador treinado (usar a API para enviar o texto para ser classificado e o serviço retorna a classe que combina com o texto); 4. Avaliar os resultados e atualizar os dados de treinamento (atualizar a base de dados treinada com base no resultado das classificações e criar e treinar o classificador usando os dados treinados atualizados).

Para a classificação das denúncias da Linha Verde, optamos por uma simplificação do método utilizado pela Módulo [Módulo, 2013], para facilitar a identificação da classificação. Reclassificamos os valores de USR para serem apresentados através de

um código de 3 cores, a saber: verde para baixa prioridade; amarelo para média prioridade; e vermelho para alta prioridade.

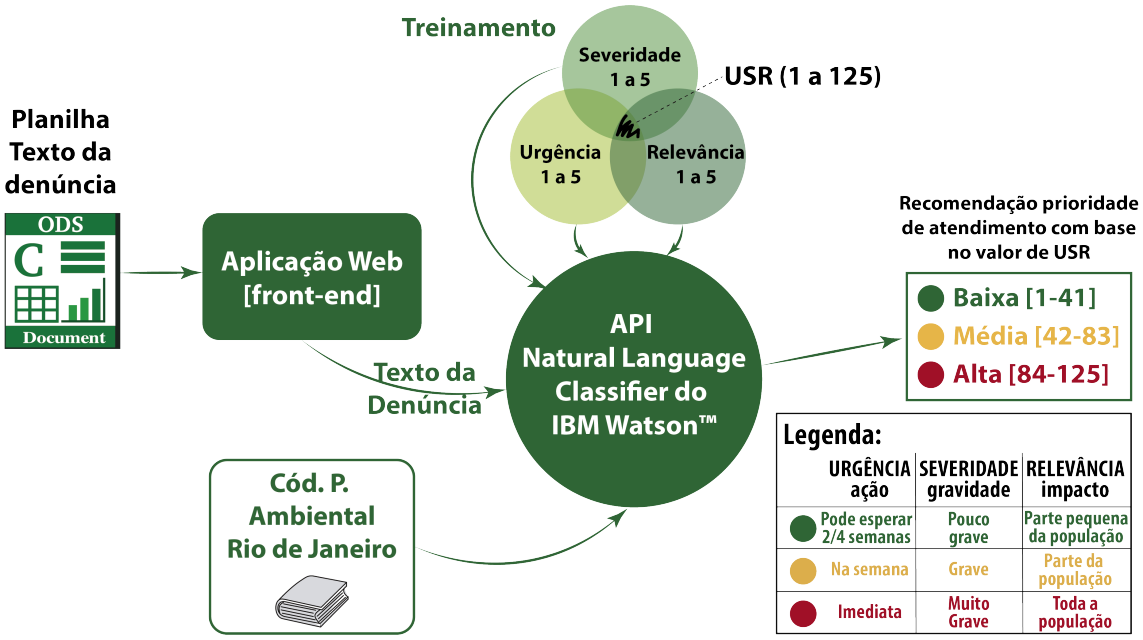


Figura 1. Diagrama da Classificação por USR

Através da utilização da metodologia de análise de riscos, encontrada na Norma ABNT NBR ISO 31000:2009 aliada ao método de classificação USR da Módulo Security for GRC e ao Código Penal Ambiental do Estado do Rio de Janeiro, conseguimos elaborar um método de classificação/recomendação de prioridade de atendimento para as denúncias, que utilizou o engenho da API Classifier do IBM Watson™ para ler e classificar o texto das denúncias. A API depois de treinada conseguiu analisar o texto, definindo os valores de urgência, de severidade e de relevância, o que permitiu a recomendação de prioridade para o atendimento das denúncias [Ricci, 2011].

Referências

Associação Brasileira de Normas Técnicas. (2009) “Norma Brasileira ABNT NBR ISO 31000:2009 - Gestão de riscos - Princípios e diretrizes”, 30 de dezembro de 2009.

DiMascio, Carmine. (2016) “Criar um classificador de língua natural que identifica spam”, IBM developerWorks. www.ibm.com/developerworks/br/library/cc-spam-classification-service-watson-nlc-bluemix-trs/index.html - visitado em 05/06/2017.

Disque Denúncia. (2017) - http://disquedenuncia.org.br - visitado em 05/06/2017.

IBM BLUEMIX Docs. (2017) “Natural Language Classifier”. https://console.ng.bluemix.net/docs/services/natural-language-classifier/natural-language-classifier-overview.html#about - visitado em 05/06/2017.

Módulo S/A ©. (2013) “Relatório de Encerramento do Projeto de Gestão Integrada de Riscos da Jornada Mundial da Juventude Rio2013 para o Comitê Organizador Local (COL)”, Documento Reservado, v. 1.0, Rio de Janeiro, Brasil.

Ricci, Francesco. (2011) “Recommender Systems Handbook”, New York: Springer, cap. 5, pp. 39-67, 145, 387-399, 645-672.

Um Método para Implantação de Boas Práticas de Gestão de Serviços de TI em Instituições de Ensino

Lucas B. Martins, Rômulo S. Sanglard, João V. Guinelli, Eliezer D. Gonçalves, Rafael E. L. Escalfoni

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET-RJ)
Av. Gov. Roberto Silveira, 1900, Prado, Nova Friburgo, RJ – Brasil

{lucasbertoloto12,romulossanglard}@gmail.com

{joao.silva, eliezer.goncalves, rafael.escalfoni}@cefet-rj.br

Abstract. *Higher investments demands a better and more disciplined quality of IT services. Properly managing resources and risks, delivering business value is one of the challenges of the IT industry. This work presents a method to adopt ITIL best practices, according to an iterative and incremental model.*

Resumo. *Os altos investimentos têm exigido uma oferta melhor e mais disciplinada dos serviços de TI. Gerenciar adequadamente os recursos e riscos, entregando valor ao negócio é um dos desafios do setor de TI. O presente trabalho apresenta um método para adoção de boas práticas do ITIL®, segundo um modelo iterativo e incremental.*

1. Introdução

As instituições de ensino têm feito grandes investimentos em tecnologia da informação como forma de aprimorar o processo de ensino-aprendizagem e prover eficiência dos mecanismos de gestão administrativa e acadêmica [Motta 2014]. O alto custo envolvido exige uma oferta melhor e mais disciplinada dos serviços de TI. Gerenciar adequadamente os recursos e os riscos inerentes à área, assegurando o alinhamento com as necessidades de negócio e entregando informação dentro de parâmetros acordados de eficiência e disponibilidade é um dos grande desafios do setor de TI, sobretudo para organizações públicas. Diante do cenário apresentado, este trabalho consiste na apresentação de um método para implementar um modelo de gestão por processos baseado em boas práticas preconizadas pelo ITIL®.

2. Referencial Teórico

O mapeamento e a documentação de processos constituem uma importante estratégia para a organização dos processos empresariais. Através destas etapas, cria-se um entendimento comum sobre diferentes setores, aumentando a compreensão dos problemas e oportunidades de negócio. [Maranhao and Macieira 2014]. A gestão de processos de negócio tem evoluído ao longo dos últimos anos sob a forma de *Business Process Management* (BPM) - uma disciplina associada a um conjunto de tecnologias que apoiam o gerenciamento de processos [Souza 2016].

2.1. Gestão de Serviços de TI

De acordo com Motta (2014), a gestão de TI é responsável pelos processos de administração dos recursos de TI, assegurando eficiência e apoiando as decisões acerca dos serviços de TI prestados. O ITIL é um arcabouço de boas práticas para a definição de modelos de gestão de serviços de TI baseado em processos escaláveis. Por conta de sua abrangência e profundidade, tornou-se um padrão mundial. O ITIL® v3 é fundamentado no Ciclo PDCA, no qual as ações são implementadas de maneira iterativa e incremental. As fases propostas pelo núcleo do ITIL® são complementares e interdependentes. Desta forma, os livros não podem ser utilizados isoladamente, mas seus processos podem ser implementados de maneira progressiva [Freitas 2013].

3. Proposta de Método

A proposta é baseada em quatro etapas complementares e cíclicas, baseada no ciclo PDCA. Na primeira etapa, de **Diagnóstico**, busca-se compreender o ambiente, a estrutura e a dinâmica da organização. Deve-se fazer uma modelagem dos processos existentes. A segunda etapa, de **Planejamento**, serve para definir as ações corretivas, os sistemas de apoio e as propostas de modelos aprimorados. A terceira etapa, de **Proposição**, é a parte de execução das ações propostas na etapa de Planejamento. Por fim, na etapa de **Monitoramento**, deve-se verificar a eficácia das ações implantadas e compreender o que pode ser melhorado nas próximas iterações.

4. Considerações Finais

O método apresentado já está sendo aplicado em uma instituição pública de ensino, tendo completado um ciclo. Nesta iteração, foi feito um diagnóstico inicial, através de entrevistas no campus. Depois, foram modelados os processos referentes aos serviços oferecidos pelo setor de TI, resultando em uma proposta inicial de gerenciamento de portfólio de serviços. Em seguida, foi escolhida a ferramenta *iTop* para apoiar a gestão de recursos de TI. Foram propostas diversas adaptações aos processos existentes, adequando-os às práticas recomendadas pelo ITIL®. Foram catalogados todos os equipamentos do campus no *iTop*, dando início aos primeiros processos relacionados à Gestão de Configuração. Em seguida, foram definidos os processos para a criação de uma Central de Serviços.

Com o objetivo de verificar a validade e a aceitação da proposta, pretende-se realizar um estudo qualitativo sobre o método. Para isto, estão sendo elaborados questionários a serem aplicados à gerência e aos demais clientes. Também está sendo analisado o impacto da ferramenta para o sucesso da proposta.

References

Freitas, M. A. S. (2013). *Fundamentos do Gerenciamento de Serviços de TI*. Brasport, Rio de Janeiro, 2 edition.

Maranhao, M. and Macieira, M. E. B. (2014). *O processo nosso de cada dia: modelagem de processos de trabalho*. Qualitymark Editora, Rio de Janeiro, 2 edition.

Motta, G. T. (2014). Serviços de tecnologia da informação: Fator de sucesso na governança e gestão das ies. In *Congresso Nacional de Excelência em Gestão*.

Souza, M. G. S. (2016). Melhoria nos processos de negócios do centro de tecnologia da informação e comunicação (ctic) da universidade federal do amazonas.

Análise de Crédito Financeiro através do classificador FAT-WiSARD

Pedro M. Xavier[1], **Alan T. L. Bacellar**[1], **Diego F. P. de Souza**[2], **Hugo C. C. Carneiro**[2], **Felipe M. G. França**[2]

[1]Escola Politécnica – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, RJ – Brazil

[2]Programa de Engenharia de Sistemas e Computação (PESC)
COPPE/UFRJ – Rio de Janeiro, RJ – Brazil

{pedromxavier, alanbacellar}@poli.ufrj.br

Abstract. *This paper presents the results obtained using a new model of Neural Network to discern between good and bad clients among the thousand records of the* ***German Credit*** *database. The instances of this collection have, each, twenty attributes about credit requesters in German banks. The information, in fact, constitues a very heterogeneous and noisy set of data, requiring new approaches to solve the problem.*

Resumo. *Este trabalho apresenta os resultados obtidos utilizando-se um novo modelo de Rede Neural para discernir entre bons e maus clientes dentre os mil registros da base de dados* ***German Credit****. As instâncias desta coleção possuem, cada uma, vinte atributos acerca de requisitantes de crédito em bancos alemães. As informações constituem, de fato, um conjunto de dados bastante heterogêneo e ruidoso, o que demanda novas abordagens para a solução do problema.*

1. Introdução

Temos como motivação deste estudo um conhecido problema de Análise de Crédito [Hofmann 1994] onde propomos um novo modelo de Rede Neural, que se mostra uma intersecção entre as Redes com pesos [McCulloch and Pitts 1943] e as Redes sem pesos [Aleksander et al. 1984]. Os dados provém do repositório *UCI* [Lichman 2013] para aprendizado de máquina.

2. Análise de Crédito e a *FAT-WiSARD*

Os dados consistem em 1000 exemplos de clientes que requisitaram crédito financeiro e que, após o período de empréstimo, foram classificados, conforme suas atitudes, como bons (700 exemplos) ou maus pagadores (300 exemplos). No entanto, a amostra se mostrou bastante heterogênea e ruidosa, além de apresentar variáveis de diferentes tipos: Discretas, contínuas, categóricas e qualitativas. Por isso, substituímos as memórias *RAM* presentes no modelo *WiSARD* original por espaços euclidianos ($\mathbb{R}^n$) onde encontram-se diversos pontos, análogos aos endereços de memória, posicionados durante o treinamento. O algoritmo consiste, então, em classificar um ponto de entrada com base nos pontos que estão presentes na sua vizinhança, delimitada por uma bola de raio r. Nas

entradas da rede foram utilizados pesos, para selecionar os atributos que serão mais relevantes para a classificação. Na saída de cada espaço temos pesos também, mais uma vez escolhendo aqueles que se destacam na classificação, sendo assim análogo ao *bleaching* [Carvalho et al. 2013].

3. Resultados

Após o treinamento da rede, comparamos o desempenho da *FAT-WiSARD* com o modelo *ClusWiSARD* [Cardoso et al. 2016] e a *Support Vector Machine* (SVM) [Haltuf 2014]. A partir dos resultados coletados estruturamos a seguinte tabela:

	FAT-WiSARD	**ClusWiSARD**	**SVM**
Accuracy	0.785	0.767	0.765
F1(Good)	0.853	0.841	0.843
F1(Bad)	0.598	0.563	0.540

4. Conclusão

A *FAT-WiSARD* demonstrou um bom resultado perante os dados utilizados. No entanto, é interessante, como atividade futura, verificar se o algoritmo comporta-se de modo estável quando apresentado a outros conjuntos de informações. Outra questão importante é o tempo de máquina demandado pelo modelo, que apesar de mais acurado do que os demais, se mostrou mais custoso. Portanto, a definição de uma nova métrica que não a euclidiana, assim como uma outra forma de determinar a vizinhança do ponto de entrada, podem ser caminhos para tornar o modelo mais competitivo quanto à sua eficiência.

Referências

Aleksander, I., Thomas, W., and Bowden, P. (1984). Wisard, a radical new step forward in image recognition. 2:120–124.

Cardoso, D. O., Carvalho, D. S., Daniel S. F. Alves, D. F. P. S., Carneiro, H. C. C., Pedreira, C. E., Lima, P. M., and França., F. M. (2016). Financial credit analysis via a clustering weightless neural classifier. *Neurocomputing*, 183:70–78.

Carvalho, D. S., Carneiro, H. C. C., França, F. M. G., and Lima., P. M. V. (2013). B-bleaching: Agile overtraining avoidance in the wisard weightless neural classifier. *ESANN 2013 proceedings*, pages 515–520.

Haltuf, M. (2014). Support vector machines for credit scoring. Master's thesis, University of Economics in Prague Faculty of Finance.

Hofmann, H. (1994). German credit data. Institut für Statistik und Ökonometrie, Universität Hamburg.

Lichman, M. (2013). *UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences* **[http://archive.ics.uci.edu/ml]**.

McCulloch, W. S. and Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. 5:115–133.

Agradecimentos

Os autores agradecem ao Prof. Dr. Felipe França, pela orientação e pelo conhecimento transmitido; assim como ao Dr. Hugo Carneiro e ao MSc. Diego Souza.

Mobile Recommendation System with Crowdsourcing and Geospatial Data

Kleyton P. Cotta[1], **Raul S. Ferreira**[1,2], **Carlos Eduardo Barbosa**[1], **Jano Moreira de Souza**[1]

[1]COPPE - Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brasil

[2]DCC - Universidade Federal Rural do Rio de Janeiro (UFRRJ)
Rio de Janeiro, Brasil

kpcotta,raulsf,eduardo,jano{@cos.ufrj.br}

Abstract. *Geospatial services have been used in several application areas. In this work, we create a mobile crowdsourcing tool aiming to assist people to publish, finding and recommend nearby services from small business. Our model is based on the concepts of Computer Supported Cooperative Work. The use of crowdsourcing for voluntary production of information is stimulated with a gamification system, which improves the data quality and keep users engaged. A recommendation engine is integrated in order to display personalized information and to keep the user interested inside the application. We show how we developed this framework and show a case study in a small business segment.*

1. Introduction

Applications that exploits, collects and disseminates geographic information, such as Google Maps, OpenStretMap and Wikimapia have attracted, for many years, as users as developers. Therefore, can be used for various purposes, such as education, entertainment, traffic and many other areas. The success of these applications relies in the crowdsourcing component. Crowdsourcing is a term that can be seen as an overlay of the terms ”wisdom of crowd” and ”outsourcing” [Muhammadi and Rabiee 2013]. According to [Howe 2006] crowdsourcing refers to a group of people that can converge to the solution of an individual problem, in which a specialist may not be able to solve.

Research conducted by the Sebrae (Brazilian service of assistance to micro and small enterprises)[1] in 2015, show that 95% of the companies in Brazil are represented by small businesses, which 44% are in commerce and 35% are in the service sector, and generates about 17 million jobs with a formal contract and reach 27% of the Gross Domestic Product, thus, we can see how important is the small businesses inside country. Nowadays, new types of smartphones and laptops have gained space in the market through the advancement of technology, powerful processors, high-resolution display, high-speed network connectivity with navigation features GPS and thus, there was a growth adoption of geospatial applications.

The lack of a tool that provides to users and small establishments individualized recommendations is a point that worth to be analyzed and the use of technology makes possible to further increase the participation of companies in the domestic market. Thus,

[1]https://www.sebrae.com.br/sites/PortalSebrae

this paper proposes the development of a generic model to create an information system that uses spatial data and posses recommendation and crowdsourcing mechanisms in order to establish data reliability to its users and suppliers, providing a template application for mobile recommendation system, which receives voluntary contributions of geographic and textual data and reviews of the quality of service from an establishment as well.

The remaining parts of this paper is organized in the following way: Section 2 discusses related work, focusing on collaborative development for mobile devices. Section 3 introduces our proposed model for creating the system. In Section 4 we apply and validate our model with a small business case of study and we finish with section 5 bringing a summary and an outlook on future work.

2. Related Work

In the early 70s, there is an increase in the fields of software and Office Automation Engineering. The purpose of these areas was to give computer support to large groups involved in projects. In 1988, Paul and Irene Greif Cashman, came up with the term Computer Supported Cooperative Work (CSCW) partly as a shorthand way of referring to a set of concerns about the support of multiple individuals working in conjunction with computer systems [Bannon and Schmidt 1989].

The crowdsourcing model for geospatial data has already been used in several real problems around the world and according to [Heipke 2010], this approach has shown that generation of content by a large number of users, for example, one of the reference projects of crowdsourcing geospatial data is OpenStreetMap[Haklay and Weber 2008], founded in 2004 by Steve Coast, with the objective of providing free access to geographic information updated worldwide. Through the data released by the own website in April 2015, it is estimated about 2.2 million registered users, 500.000 taxpayers, among these 25.000 are active. According to this paper, the quality of OpenStreetMap data obtained 80% coverage and a geometric prediction of 6 meters to the main roads in the London area in relation to the Ordnance Survey[Haklay 2010].

Another example is Wikimapia[Koriakine and Saveliev 2008], that uses Google Maps technology and allows users to delimit areas of interest and link them to descriptions and comments, allowing users to create and integrate descriptions with Wikipedia links. Another mechanism to highlight is the access policy to curb acts of vandalism, which in turn is associated with gamefication elements aiming to encourage users through scoring and ranking. According to that ranking you can have access to functions ranging from the option of adding photos to permissions for administrative privileges on the site. Another recent real example of use of geospatial data and crowdsourcing information can be seen at [Ferreira et al. 2017], which users can send their complaints and ratings about the public transport. Their locations, complaints and rankings are sent to transport supervision agencies and are stored inside a public database, which are accessible by population and researchers through an open API.

3. Framework for Web System

[Alarcon et al. 2012] presents an orientation tool to strengthen mobile collaboration and system development activities and we had its work as a basis for the development of

software to create eight categories divided into three phases: Design, analysis and architectural design. We were be able to identify a set of non-functional relevant requirements and design constraints in order to obtain a contextual application. This model is able to collect geographic data on a voluntary basis in which to assess the reliability of the contributions, providing different filtering mechanisms aimed to receive voluntary contributions of geographic data and reviews of the quality of service of an establishment. The modelling and physical schema of the database system contains the metadata that guarantee the operation of the application. Figure 1 shows schematically the operating motion to the tool with these characteristics.

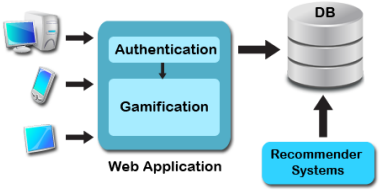


Figure 1. System Model.

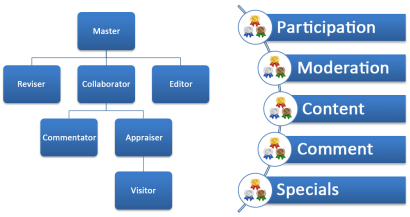
According to the model it is possible that the developer create your web and mobile system using the concepts of crowdsourcing with the data quality mechanism integrating a recommendation system. Created the structure, the developer shall be free to allow access for others to contribute. The objective is to allow access through mobile phone and computers and also streamline the application development process and systems. Besides, following the example of successful applications like Foursquare[2] and Stack Overflow[3], the idea of motivating users via game elements is generating applications in many areas, among them, finance[Deterding et al. 2011a] and academia[Huotari and Hamari 2011].

3.1. Gamification

By definition, gamification is used to represent the use of game elements to enhance the experience and user engagement in any service or application other than the context of games [Deterding et al. 2011b]. The rewards are defined as follows: Register in the app, evaluate an establishment, share, add content, moderate requests, comment and adding content, where the user can add +20, +20, +5, +2, +5, +1/-1 and +1/-1 points respectively. Figure 2(a) shows the dynamic progression of gamification system and Figure 2(b) provides the possibility to get the medals, which are divided into categories: gold, silver and bronze. Through the votes carried by users in content and comments added by a particular user can create a quality control for these data collected, thus, inhibit a member use the system inappropriately.

We use two approaches as alternatives to ensure data quality: Crowdsourcing and social approach, which aims to employ gamification to enhance this filtering data. We performed social approach based on a hierarchy of trusted individuals that act as moderators. Many studies have shown that the voluntary contributions of individuals follows a frequency distribution with a long tail with some individuals making a large number of contributions and a large number of individuals making only one or few contributions.

[2]https://foursquare.com/
[3]https://stackoverflow.com/



(a) Progression Dynamics. (b) Distribution Awards.

Figure 2. Gamification elements.

3.2. Recommender Systems

According to [Adomavicius and Tuzhilin 2005], Recommender Systems are techniques and tools used to suggest personalized items based on the interests of users. In a common system, usually the users provide the recommendations, this captured information is used by the system to present them to the groups of individuals considered potential interested for this type of recommendation. Recommender systems are designed to filter information according to the profile of interests of the users and thus, recommend items that meet the expectations and needs of users and are generally used in one of the following three aforementioned information filtering techniques: Content-based filtering, collaborative filtering, also known as social filtering and hybrid filtering.

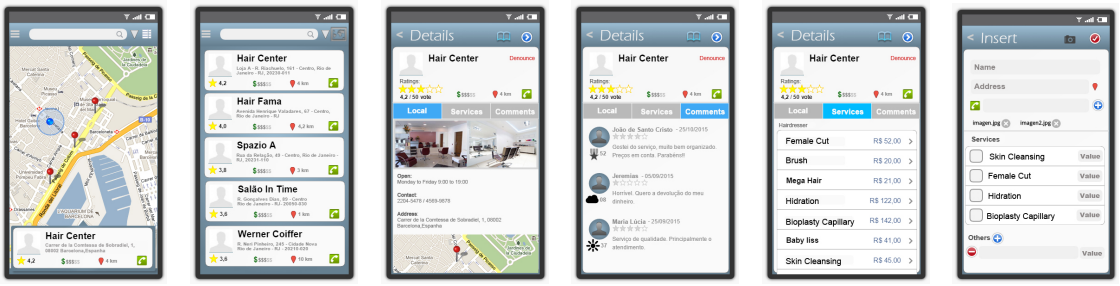
For this work we chose the collaborative-filtering approach because this method analyze large amounts of information about users' preferences and predict the preferences of similar users to recommend items. Thus, it is possible to make an accurate prediction of a user's preferences and deliver items recommendations without any need for a detailed analysis of item characteristics. The technique relies on analysis of common preference in a group of people. The essence of this technique is the exchange of experiences among people who have common interests and have similar choices for items. Besides, this technique constitutes one of the most popular recommendation techniques being used in many existing systems on the Internet [Schafer et al. 2001]. In collaborative systems, the essence is the exchange of experiences among people who have common interests. In these systems, the items are filtered based on evaluations by users.

4. Experiments

In this section we present an example using this model in the area of aesthetics and beauty. A need in this area was realized due to a number of existing establishments that do not use so much publicity and that most of the disclosure of the establishment is carried out by word of mouth, so we use the halls and aesthetic houses as inputs to this model. Below we present the key interfaces in the system, using the data mentioned above and showing how would be the implementation of the model for the mobile environment. Figure 3(a) shows how the pursuit of establishments are presented across the map. In that case, are made a new query through the point of displacement on the map, since the figure 3(b) shows the search result by a list. Both ways have filters to refine the results and address bar to perform a search.

As seen in the previous figures, the result of searches screens also feature qualifying, the distance in relation to the user and the average price of the establishments. An-

other feature of this screen is that is possible to receive establishments recommendations and make connections with the same categories establishments. The system interface is responsible for detailing the establishment. It shows the photo, the location, the score, the services offered, comments, and average prices. Figure 3(c) shows the information about the location of the property while Figure 3(d) shows the user feedback with their respective ratings already Figure 3(e) shows the services and their values and finally Figure 3(f) allows the employee to add values and services Of the establishment.



(a) Results Map. (b) Result List. (c) Local Tab. (d) Comments Tab. (e) Details Tab. (f) Insert Tab.

Figure 3. Interface of the establishment of the information.

To evaluating the model, the system carry out the evaluations of the establishments were chosen five subjects that best represents the nature of a rating, varying from 1 to 5 (higher is better) to identify better the differences between properties: Attendance; Location; Hygiene; Cost-benefit; Product quality. The calculation of the final grade given to the user establishment is the result of the average grade of the subjects. However, the rating of the establishment is developed in the frequency histogram of the ratings of the users. The ratings of the establishments along with the users of the profile is used as input to the recommendation system model. Which in turn hold their predictions thus, individualized information is passed to each user of the system, being added to the results obtained by the search or the news tab and promotions.

5. Conclusions

In this paper, we introduced the importance of this system to assist people in small business activities and through the use of the concepts of crowdsourcing applied to geospatial data. It was possible to develop effective supporting tool for group work within a system. The concepts derived from the theoretical framework and related work obtained an important role in the creation and design of the proposed model, so it was possible to meet the requirements encountered during the development and study of this model.

Therefore, another feature was the ability to keep users engaged in the project using gamification. Using the recommendation system in our model facilitated the view of information from users perspective and further incremented the loyalty of the user inside the system. Thus, this model can be applied to various business niches, such as in the area of automotive, education and food branch. Its generalization capability facilitates the development and adaptation of this system to suit according to the selected market segment. With the increasing technological advancement especially in mobile computing, the trend is to further intensify the use of systems based on crowdsourcing. However, we see the

need to determine best approaches and algorithms to be applied in the recommendation system. These in-depth analyzes with comparison of these techniques are to be performed in future work.

References

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749.

Alarcon, R., Guerrero, L. A., Ochoa, S. F., and Pino, J. A. (2012). Analysis and design of mobile collaborative applications using contextual elements. *Computing and Informatics*, 25(6):469–496.

Bannon, L. J. and Schmidt, K. (1989). Cscw-four characters in search of a context. *DAIMI Report Series*, 18(289).

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011a). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15. ACM.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., and Dixon, D. (2011b). Gamification. using game-design elements in non-gaming contexts. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 2425–2428. ACM.

Ferreira, R. S., Prata, R., Barbosa, C. E., de Souza, J. M., Mororo, V., and Cotta, K. P. (2017). Transreport: collaborative supervision of the public transportation. In *Proceedings of the Symposium on Applied Computing*, pages 1808–1813. ACM.

Haklay, M. (2010). How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning. B, Planning & design*, 37(4):682.

Haklay, M. and Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18.

Heipke, C. (2010). Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):550–557.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Huotari, K. and Hamari, J. (2011). Gamification" from the perspective of service marketing. In *Proc. CHI 2011 Workshop Gamification*.

Koriakine, A. and Saveliev, E. (2008). Wikimapia. *Online: wikimapia. org*.

Muhammadi, J. and Rabiee, H. R. (2013). Crowd computing: a survey. *arXiv preprint arXiv:1301.2774*.

Schafer, J. B., Konstan, J. A., and Riedl, J. (2001). E-commerce recommendation applications. In *Applications of Data Mining to Electronic Commerce*, pages 115–153. Springer.

Text Mining em documentos de patentes do USPTO: Um Estudo de caso usando o RapidMiner

João Marcos de Rezende[1], **Karin Satie Komati**[1], **Leandro Colombi Resendo**[1]

[1] Programa de Pós-Graduação em Engenharia de Controle e Automação (ProPECAut)
Instituto Federal do Espírito Santo - Campus Serra
Rodovia ES 010 – Km 6,5 – Manguinhos – Serra – ES – Brasil

Abstract. *There are several tools of Data Mining, many of them already have features of Text Mining to analyze large amount of textual documents. The proposal of the work was based on the use of the Text Mining technique for the collection of keywords from patent documents and the application of the k-Means algorithm for clustered grouping of these documents in RapidMiner software. In this way, present itself the step-by-step from the data collection to the presentation of the results. The performance of RapidMiner software was investigated at the USPTO office's patent base from january/2010 to march/2015.*

Resumo. *Há várias ferramentas de Data Mining, muitas delas já possuem funcionalidades de Text Mining para analisar grande quantidade de documentos textuais. A proposta do trabalho baseou-se na utilização da técnica de Text Mining para a coleta de palavras-chave de documentos de patentes e, aplicação do algoritmo k-Means para o agrupamento de forma clusterizada destes documentos no software RapidMiner. Desta forma, apresenta-se o passo-a-passo desde a coleta dos dados até a apresentação dos resultados. Investigou-se a performance do software RapidMiner na base de patentes do escritório USPTO no período de janeiro/2010 à março/2015.*

1. Introdução

O Instituto de Nacional da Propriedade Intelectual [INPI 2017] define patente como sendo um título de propriedade temporária sobre uma invenção ou modelo de utilidade, conferido pelo Estado aos inventores ou outras pessoas físicas/jurídicas detentoras de direitos sobre a criação. Assim, o inventor ou o detentor da patente tem o direito de impedir que terceiros, sem o seu consentimento, possam produzir, usar, vender ou importar o objeto de sua patente. Em contrapartida, o inventor é obrigado a revelar detalhadamente todo o conteúdo técnico da matéria resguardada pela patente.

Antes de depositar o pedido de patente, é recomendável que se faça primeiro uma busca para saber se há registro de produto/processo igual ou semelhante. Uma patente registrada no INPI, em geral, é válida somente em território nacional. Caso queira que a patente seja válida em outros países, então é preciso depositar um pedido equivalente nesses outros países onde se deseja obter a patente. O pedido, depositado no Brasil, deverá ser traduzido para o idioma do país onde se deseja depositar e deverá ser nomeado um procurador para representar a empresa naquele país. O procedimento de depósito em diferentes países pode ser simplificado, ao usar o Tratado de Cooperação de Patentes [WIPO 2017], no qual o INPI atua como escritório receptor no Brasil.

No entanto, a busca de patentes pode ser uma tarefa árdua. Além da quantidade de patentes registradas ser vasta, há sinônimos e homônimos nas terminologias técnicas e jurídicas. Portanto, as informações sobre patentes precisam ser transformadas em algo mais simples e mais fácil de entender [Kim et al. 2008, Gusberti and Schunke 2016]. Ademais da dificuldade na análise das patentes, os escritórios de depósitos de patentes [INPI 2017, USPTO 2017, WIPO 2017, EPO 2017] que disponibilizam o *donwload* gratuito dos documentos, só auxiliam na busca das patentes através de palavras-chave. No entanto, não oferecem ferramentas de análise para extração de conhecimentos.

O termo *Patinformatics* [Trippe 2003] compreende o uso ou o desenvolvimento de ferramentas automatizadas para revelar a inteligência contida em um conjunto de patentes através de técnicas como visualização, análise de citações, análise de tendências e aplicações de técnicas de *Text Mining* e *Data Mining*. O *Data Mining* (Mineração de Dados) procura descobrir padrões emergentes de banco de dados estruturados, já o *Text Mining* (Mineração de Textos) extrai conhecimento útil de dados não-estruturados ou semi-estruturados [Barion and Lago 2015].

As ferramentas para trabalhar com *Data Mining* podem ser classificadas em 3 categorias: ferramentas de *Data Mining* tradicionais, *dashboards* e ferramentas *Text Mining*. Os programas tradicionais monitoram tendências de dados e podem capturar informações que residam até fora de um banco de dados. Os *dashboards* refletem de forma gráfica as mudanças e atualizações dos dados. As ferramentas *Text Mining* tem este nome pela habilidade de minerar dados em diferentes fontes de texto. Alguns *softwares* utilizados para este fim são: Weka, RapidMiner, Tanagra, DBMiner, Witness Miner e Orange [Ramamohan et al. 2012, Boscarioli et al. 2014].

Organizar dados em grupos é um dos modos fundamentais extração de informação. A análise de cluster é o estudo formal de métodos e algoritmos para agrupar objetos com características similares. Um dos algoritmos de agrupamento mais populares e simples é o *k-Means*, que foi publicado pela primeira vez em 1955. Mesmo que existam muitos outros algoritmos de agrupamento publicados, o *k-Means* é ainda amplamente utilizado [Jain 2010].

Este trabalho apresenta uma pesquisa exploratória e aplicada visando avaliar a performance do *software* RapidMiner em base de dados de patentes. A proposta baseia-se na utilização da técnica de *Text Mining* para a coleta de palavras-chave dos documentos de patentes e, fazer o uso do algoritmo *k-Means* para o agrupamento de forma clusterizada destes documentos. A escolha do RapidMiner foi baseada na facilidade de usabilidade [Viterbo et al. 2016].

Na próxima seção apresentam-se os conceitos relacionados a *Data Mining* e alguns trabalhos que utilizam esta técnica para análise de patentes. Na Seção 3 descreve-se a metodologia utilizada para coletar e armazenar os dados. A Seção 4 apresenta a análise dos dados coletados e, por fim, a Seção 5 indicam-se as principais contribuições e trabalhos futuros deste artigo.

2. Referencial Teórico

Existem dois principais sistemas de classificação de patentes: o que é utilizado pelo *United States Patent and Trade Office* (USPTO) e o *International Patent Classification*

(IPC), criado pela Organização Mundial de Propriedade Intelectual (WIPO - *World Intellectual Property Organization*). Então, embora existam diversos escritórios de patentes no mundo, as patentes são classificadas por áreas tecnológicas conforme definido pelo IPC, mesmo que internamente em cada escritório haja outra classificação. O USPTO é considerado o escritório mais valioso, porque o mercado norte americano é líder em diversas tecnologias. Com isso, diversas empresas tendem a depositar suas patentes nos Estados Unidos para assegurar a propriedade intelectual de seus inventos, em grande escala de mercado [Leydesdorff et al. 2014].

Um documento de patente contém vários atributos que podem ser separados em dois grupos: itens estruturados e não-estruturados. Os itens estruturados são uniformes em semântica e em formato como número de patente, data de depósito, autor, entre outros. Por outro lado, os itens não estruturados são textos livres e muito diferentes, tanto em comprimento e quanto em conteúdo, tais como os dados de reivindicações, resumos ou descrições da invenção [Kim et al. 2008].

Os documentos de patentes nem sempre podem ser analisados diretamente devido aos itens não-estruturados, em linguagem natural. A Mineração de Textos é o processo para descobrir padrões a partir de dados textuais e tem sido amplamente utilizado na recuperação de informações nos documentos de propriedade intelectual [Lee et al. 2008]. Existem diversas técnicas de Mineração de Textos como: segmentação de texto, extração de resumo, seleção de recursos, associação de termos, clusterização, identificação de tópicos e mapeamento de informações [Tseng et al. 2007]. Os mesmos autores definem que, de uma forma geral, a metodologia para trabalhar com *text mining* em documentos de patentes, deve seguir os seguintes passos:

1. Processamento dos documentos: coletar os dados, analisar e segmentar, resumir o texto e selecionar trechos para substituição;
2. Indexação: extração de frases ou palavras-chave, análise morfológica, filtragem de *stopwords*, associação de termos e clusterização;
3. Clusterização: seleção de termos, categorização e clusterização dos documentos, geração de títulos dos clusters e mapeamento das categorias e
4. Mapeamento: mapeamento de tendências, consultas do mapeamento, agregações e análise de profundidade.

3. Estudo de Caso

Esta seção descreve as características da ferramenta investigada neste trabalho e, as seções que seguem apresentam as etapas principais ao longo desta pesquisa.

3.1. Ferramenta

O RapidMiner é um software para análise preditiva e aprendizagem de máquina. Uma de suas características principais é a execução de comandos sem a necessidade de desenvolvimento de código, apenas utilizando a interface gráfica [RAPIDMINER 2014]. O usuário inclui elementos gráficos que significam uma operação em questão e vai unindo os elementos, de tal forma a construir um fluxo de execução intuitivo.

De acordo com [Rexer et al. 2015], o RapidMiner começou a ser mencionado (entre os sistemas de *Data Science* que disponibilizam versões *free*) com maior significância

a partir de 2015. Ainda em relação à pesquisa destes autores, entre os usuários que utilizaram o software, 32% disseram estar extremamente satisfeitos e 59% satisfeitos.

Neste trabalho foi utilizado o *plugin Text Processing* que, por sua vez, possui diversas funções para tratar textos conforme as técnicas mencionadas na Seção 2. Há outros *plugins* que fazem tratamentos de textos, como exemplo o *Information Extraction*. Porém, este último, possui funções mais voltadas para extração e classificação das palavras, de acordo com o conceito de *Named Entity Recognition* [Jungermann 2009].

3.2. Coleta de Dados

A coleta dos dados de patentes foi efetuada através do *download* de arquivos XML, disponibilizado pelo Google [GOOGLE 2012], do escritório de patentes dos Estados Unidos, o USPTO. Os arquivos coletados foram de janeiro/2010 à março/2015, de patentes que já foram publicadas e concedidas ao depositador da patente.

Estes arquivos são formatados seguindo o layout *U.S. Patent Grant Data/XML v4.3*. Porém o intuito foi a análise das informações textuais dos títulos e dos resumos das patentes. Além das tags já previstas no documento de *layout*, haviam tags HTML de formatação, que foram removidas. Assim, foi desenvolvido um programa para extrair apenas os dados necessários para análise e, posteriormente, a criação de um novo arquivo, já formatado para importação para o banco de dados. Os dados extraídos foram: número do documento, data de publicação, data de aplicação, seção, classe, sub-classe, grupo principal, sub-grupo, título e resumo.

Os novos arquivos foram importados em um banco de dados MySQL, facilitando, assim, o manuseio das informações através de consultas SQL para selecionar os dados utilizados pelo RapidMiner.

3.3. Elaboração das Análises

O RapidMiner foi configurado para realizar alguns processos de *Text Mining* para prover os resultados obtidos, conforme mostrado na Figura 1. A primeira caixa representa a conexão com o banco de dados, a segunda representa o tratamento dos textos e, por último, a aplicação do algoritmo de clusterização, o *k-Means*.

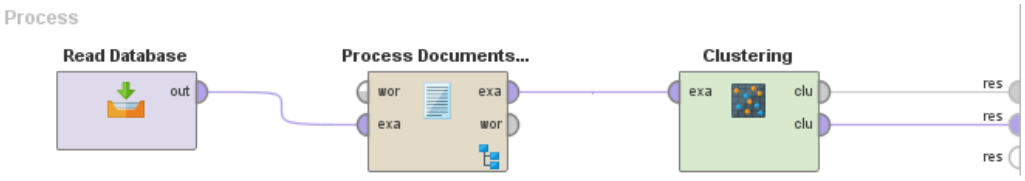


Figura 1. Processo no RapidMiner

Após a leitura dos dados, estes passaram por transformações, conforme o fluxo apresentado na Figura 2. As funções estão internas na segunda caixa da Figura 1 e são disponibilizadas pelo *plugin Text Processing*, conforme mencionado na seção 3.1.

- os textos foram colocados em letras minúsculas. Na Figura 2 é o elemento *Transform Cases*;
- aplicação do filtro de *stopwords* (palavras que são consideradas irrelevantes ao contexto geral de um conjunto de textos) do idioma inglês, como por exemplo: *if*, *an*, *to*, dentre outros. Na Figura 2 é o elemento *Filter Stopwords*;

- tokenização dos textos por patente, ou seja, um token representa um grupo de palavras de uma única patente. Na Figura 2 é o elemento *Tokenize*;
- as palavras consideradas para análise foram parametrizadas para ter no mínimo quatro e no máximo vinte e cinco caracteres. Bem como, a aplicação de expressões regulares para retirar símbolos e números. Na Figura 2 são os elementos *Filter Tokens*.

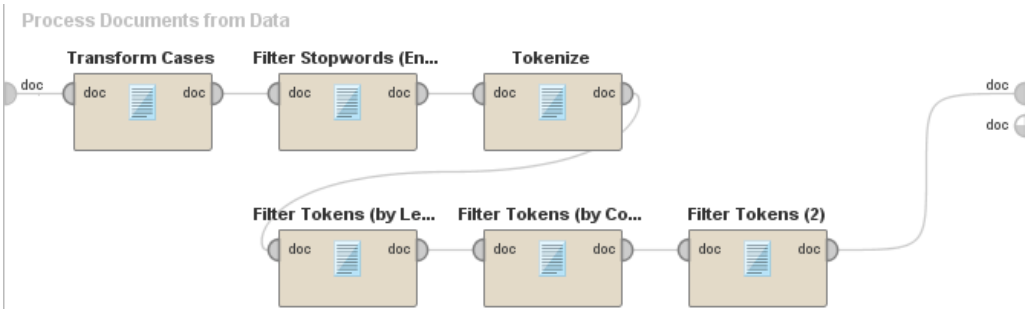


Figura 2. Funções do plugin *Text Processing*

Por fim, houve a inserção do algoritmo *k-Means* (terceira caixa da Figura 1) para realizar a clusterização das palavras, de forma a mostrar como os dados são agrupados pelo conjunto de palavras.

O algoritmo *k-Means* tem o seguinte funcionamento: o número k de grupos que se deseja encontrar precisa ser informado de antes da execução. Em seguida, k pontos são escolhidos aleatoriamente para representar os centróides dos grupos, com isso, um conjunto de elementos é particionado de forma que cada elemento é atribuído à um grupo, de centróide mais próximo, de acordo com a distância euclidiana comum. A cada iteração do algoritmo, os k centróides, ou "médias"(e daí vem o nome *means*), são recalculados de acordo com os elementos presentes no grupo e em seguida todos os elementos são realocados para a partição cujo o novo centróide se encontra mais próximo [Costa et al. 2013].

4. Resultados

A seguir, serão apresentados alguns experimentos realizados e seus respectivos resultados. Os testes foram executados em computador com as seguintes configurações: sistema operacional Windows 10 64bits, processador Intel i7-6850k 3.60GHz e memória RAM de 64GB.

4.1. Resultado Geral dos Resumos das Patentes

Para a realização deste experimento, foi utilizado como entrada de dados todos os 1.346.684 registros de patentes do banco de dados, conforme descrito na Tabela 1. Neste experimento, os dados não estavam balanceados, ou seja, não havia a mesma quantidade de patentes para cada Seção (neste caso, Seção refere-se à classificação da área tecnológica definida pelo IPC).

O processo foi executado em 1 dia 13h34m24s. A Tabela 2 mostra o resultado do algoritmo, onde é possível observar que os clusters tem quantidades de itens diferentes do que foi utilizado como entrada de dados na Tabela 1. Os clusteres da Tabela 2 não referem-se na ordem respectiva de cada Seção da Tabela 1. O agrupamento dos itens em

Tabela 1. Quantidade de Patentes por Seção

Seção	Nº Patentes	Nº Caracteres
A - HUMAN NECESSITIES	182376	112757518
B - PERFORMING OPERATIONS; TRANSPORTING	160413	110543913
C - CHEMISTRY; METALLURGY	93263	53943777
D - TEXTILES; PAPER	5896	3807953
E - FIXED CONSTRUCTIONS	27418	18855911
F - MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING	76422	53390769
G - PHYSICS	439413	316584021
H - ELECTRICITY	361483	258953535
Total	1346684	928837397

cada cluster é realizado de acordo com a frequência de ocorrência de cada palavra, ou seja, para a clusterização identificar exatamente as patentes que pertencem à mesma área tecnológica, o número de patentes de cada cluster da Tabela 2, deveria ser igual ao número de patentes de alguma Seção da Tabela 1.

Tabela 2. Clusterização por Seção desbalanceada

Cluster	Nº Patentes
0	183152
1	100505
2	89728
3	62340
4	119052
5	101908
6	583555
7	106444

Este experimento serviu como base de análise para executar os próximos testes, pois o tempo de resposta pode ser uma restrição até como decisão de escolha deste software como ferramenta de mineração de dados.

4.2. Resultados dos Resumos de Patentes com base Reduzida

Neste caso, foi utilizado como entrada de dados 5.896 registros das Seções de A a H. A Tabela 3 exibe a quantidade de caracteres que foram analisados por Seção. A escolha deste número de registros foi pelo fato de que a Seção C só contém essa quantidade de patentes.

Tabela 3. Quantidade de Caracteres dos Resumos por Seção

Seção	Nº Patentes	Nº Caracteres
A	5896	3661127
B	5896	4079067
C	5896	3484588
D	5896	3807953
E	5896	4079122
F	5896	4152597
G	5896	4272203
H	5896	4215426

Com os dados balanceados, este processo foi executado em 1h19m39s. A Tabela 4 apresenta os resultados do segundo experimento, onde também é possível analisar que os clusters tem quantidades de itens diferentes do que foi utilizado como entrada de dados na Tabela 3.

Tabela 4. Clusterização por Seção balanceada

Cluster	Nº Patentes
0	14326
1	3596
2	1910
3	4407
4	3101
5	5785
6	9681
7	4362

5. Conclusão

Este trabalho apresentou uma pesquisa com intuito de avaliar a performance do *software* RapidMiner em base de dados de patentes. A proposta baseou-se na utilização da técnica de *Text Mining* para a coleta de palavras-chave dos documentos de patentes e, aplicação do algoritmo *k-Means* para o agrupamento de forma clusterizada destes documentos.

Com base nos experimentos, pode-se concluir que, embora o *RapidMiner* apresente facilidade no uso e configuração dos processos, ainda se faz necessário ajustes nas funções utilizadas para trazer melhores resultados de classificação. É provável que ainda existam *outliers* que destoam os agrupamentos realizados pela ferramenta.

Em relação à performance em termos de tempo, foi visto através do experimento da Seção 4.2 que, ao reduzir a quantidade de registros à praticamente 3,5 % da quantidade total, o tempo de execução manteve-se proporcional. Ou seja, o fato dos dados estarem balanceados ou não, não interferiu no tempo de execução.

Para trabalhos futuros, indica-se a aplicação dos mesmos testes em outros programas de *Text Mining*, para avaliar a performance entre os mesmos. Outro ponto, é a utilização da função de *Stemming* (método para redução de um termo ao seu radical, removendo as desinências, afixos e vogais temáticas, debtre outros) para analisar a redução de *outliers* e, consequentemente, os resultados da clusterização.

Referências

Barion, E. C. N. and Lago, D. (2015). Mineração de textos. *Revista de Ciências Exatas e Tecnologia*, 3(3):123–140.

Boscarioli, C., Viterbo, J., and Teixeira, M. F. (2014). Avaliação de aspectos de usabilidade em ferramentas para mineração de dados. In *Anais da I Escola Regional de Sistemas de Informação do Rio de Janeiro*, volume 1, pages 107–114.

Costa, E., Baker, R. S., Amorim, L., Magalhães, J., and Marinho, T. (2013). Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, 1(1):1–29.

EPO (2017). European patent office. `http://www.epo.org`. Acesso em 02 de agosto de 2017.

GOOGLE (2012). Uspto bulk downloads: Patent grant full text. `https://www.google.com/googlebooks/uspto-patents-grants-text.html`. Acesso em 06 de agosto de 2017.

Gusberti, T. D. H. and Schunke, M. A. (2016). Cluster binário e mineração de patentes na inteligência de negócios para ofertantes de tecnologia. In *Anais do XVLIII Simpósio Brasileiro de Pesquisa Operacional (SBPO)*, pages 542–553.

INPI (2017). Perguntas frequentes - patente. `http://www.inpi.gov.br/servicos/perguntas-frequentes-paginas-internas/perguntas-frequentes-patente`. Acesso em 02 de agosto de 2017.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Jungermann, F. (2009). Information extraction with rapidminer. In *Proceedings of the GSCL Symposium'Sprachtechnologie und eHumanities*, pages 50–61.

Kim, Y. G., Suh, J. H., and Park, S. C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, 34(3):1804–1812.

Lee, S., Lee, S., Seol, H., and Park, Y. (2008). Using patent information for designing new product and technology: keyword based technology roadmapping. *R&d Management*, 38(2):169–188.

Leydesdorff, L., Kushnir, D., and Rafols, I. (2014). Interactive overlay maps for us patent (uspto) data based on international patent classification (ipc). *Scientometrics*, 98(3):1583–1599.

Ramamohan, Y., Vasantharao, K., Chakravarti, C. K., and Ratnam, A. (2012). A study of data mining tools in knowledge discovery process. *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, pages 2231–2307.

RAPIDMINER (2014). Rapidminer studio manual. `https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf`. Acesso em 06 de agosto de 2017.

Rexer, K., Gearan, P., and Allen, H. (2015). Data science survey. `http://www.rexeranalytics.com/assets/rexer_analytics_2015_data_miner_survey_summary_report.pdf`. Acesso em 11 de setembro de 2017.

Trippe, A. J. (2003). Patinformatics: Tasks to tools. *World Patent Information*, 25(3):211–221.

Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216–1247.

USPTO (2017). United states patent and trademark office. `https://www.uspto.gov`. Acesso em 02 de agosto de 2017.

Viterbo, J., Boscarioli, C., Bernardini, F. C., and Teixeira, M. F. (2016). Avaliação de ferramentas de apoio ao ensino de técnicas de mineração de dados em cursos de graduação. In *24 Workshop sobre Educação em Computação (WEI 2016)*, pages 2006–2015.

WIPO (2017). World intellectual property organization. `http://www.wipo.int/portal/en/index.html`. Acesso em 02 de agosto de 2017.

Desaparecidos RJ - Um Sistema de Informação Para Apoio à Busca de Pessoas Desaparecidas no Estado do Rio de Janeiro

Tadeu Moreira de Classe[1], Renata Mendes de Araujo[1], Vinícius Rodrigues Lima[1], Humberto Amaro Garcia Ferreira[2]

[1]Grupo de Pesquisa e Inovação em CiberDemocracia (CIBERDEM) - Programa de Pós-Graduação em Informática - Universidade Federal do Estado do Rio de Janeiro (PPGI - UNIRIO)

[2]Setor de Busca Eletrônica - Delegacia de Descoberta de Paradeiros (DDPA) - Polícia Civil do Estado do Rio de Janeiro (PC - RJ)

{tadeu.classe, renata.araujo}@uniriotec.br, {vini.rodrigues.90, betopersonal}@gmail.com

Abstract. *The development of technological solutions that brings together citizens and public services is extremely important to improve digital democracy. This paper describes a technological solution to support the processes performed by the Civil Police of the State of Rio de Janeiro aiming at identifying missing people. It is expected that the application provided will bring more agility to the process as well as more citizen collaboration in the identification of missing people in the state.*

Resumo. *O desenvolvimento de soluções tecnológicas que aproximem a sociedade dos processos realizados pelas organizações públicas é de suma importância para o desenvolvimento da democracia digital. Este trabalho tem o objetivo de apresentar uma solução tecnológica para apoio aos processos executados pela Polícia Civil do Estado do Rio de Janeiro na busca de pessoas desaparecidos no estado. A expectativa é que esta solução agilize os processos e possibilite uma maior colaboração dos cidadãos para identificação e busca de pessoas desaparecidas.*

1. Introdução

A democracia digital (também conhecida na literatura como democracia eletrônica ou ainda e-democracia) entende a internet como agente de transformação dos processos políticos e de prestação de serviços públicos tradicionais (Vedel, 2006). Silva (2005) (Araujo et al., 2011). Acredita-se que o uso de TICs (Tecnologias de Informação e Comunicação) promova o engajamento de cidadãos na construção de políticas, na participação social, e na melhoria da prestação de serviços públicos. Neste tema, o grupo de pesquisa e inovação em Ciberdemocracia[1] (CIBERDEM) da UNIRIO (Universidade Federal do Estado do Rio de Janeiro) desenvolve pesquisas na área de governo, participação e democracia digital, como forma de transformação social do setor público e organizações de maneira geral. No contexto das pesquisas do

[1] CIBERDEM - http://ciberdem.uniriotec.br/wordpress/

CIBERDEM, têm sido construídas soluções para promover a aproximação entre governo e sociedade, algumas delas em parceria com entidades públicas nacionais.

Neste artigo apresentamos uma destas soluções, voltada à aproximação dos cidadãos ao processo de busca de pessoas desaparecidas no Estado do Rio de Janeiro. A Delegacia de Descoberta de Paradeiros (DPPA) é o setor da Polícia Civil do Estado do Rio de Janeiro especializado na descoberta de paradeiros de pessoas desaparecidas. Dentre os desafios da DPPA, está como ampliar a disseminação de informação a respeito de pessoas desaparecidas no estado e como atrair e engajar pessoas no processo de identificação de desaparecidos (vide Seção 2). Apresentamos o desenvolvimento de uma solução tecnológica - um sistema web e um aplicativo móvel - que apoia os processos executados pela DPPA, visando agilizar a solicitação de informações para identificação de desaparecidos feita por terceiros (Seção 3). O processo de desenvolvimento da solução partiu da modelagem e análise do processo de busca de desaparecidos da Polícia Civil (Classe, 2017), a partir do qual, um conjunto de funcionalidades foi identificado e implementado (Seção 3). Avaliações preliminares de uso do aplicativo foram realizadas e seus resultados são apresentados no artigo (Seção 4). Na Seção 5 concluímos a apresentação do trabalho com implicações do desenvolvimento e trabalhos futuros.

2. Desaparecimento de Pessoas no Estado do Rio de Janeiro

O desaparecimento de algum ente ou pessoa querida é uma situação que pode acontecer com qualquer pessoa, independente de contexto e da classe social. Somente no Estado do Rio de Janeiro, e segundo os dados publicados em julho de 2017 pelo Instituto de Segurança Pública do Estado do Rio de Janeiro (ISP/RJ), no período de 15 anos (2002 a 2017), foram comunicados à Polícia Civil, um número aproximado de 33 mil ocorrências de desaparecimento (Figura 1)(ISP, 2017)(Grandin e Coelho, 2017).

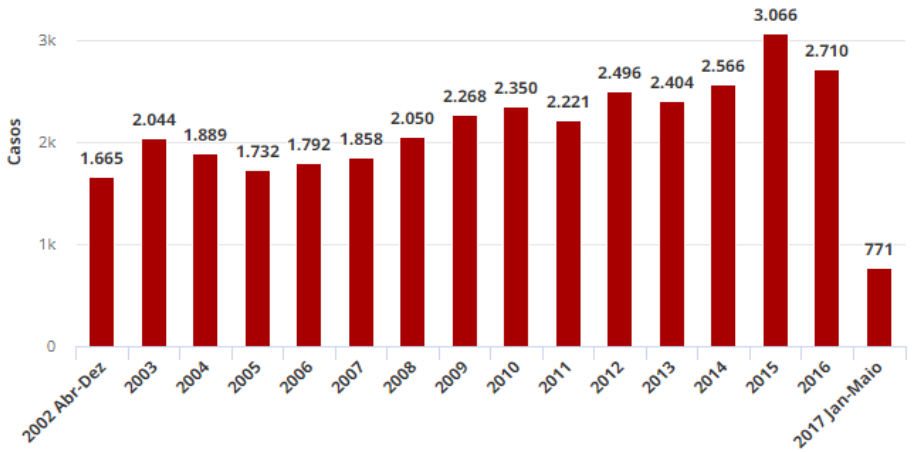


Figura 1. Registro de desaparecimento no estado do Rio de Janeiro entre 2002 e 2017 (ISP, 2017)(Grandin e Coelho, 2017).

Em setembro de 2014, foi inaugurada pelo Governo do Estado do Rio de Janeiro a DDPA[2] (Delegacia de Descoberta de Paradeiros), parte da Polícia Civil (Polícia Civil, 2014). O objetivo da delegacia é a busca e investigação de casos de desaparecimentos no estado. A criação da delegacia trouxe também a implantação do Disque-

[2] DDPA - http://www.policiacivil.rj.gov.br/exibir.asp?id=19602

Desaparecidos[3], uma forma de contribuir com informações sobre desaparecimentos de forma anônima, em funcionamento durante 24 horas ao dia, e ao alcance de todos os cidadãos. Além disso, a delegacia também oferece suporte especializado aos familiares dos desaparecidos.

Atualmente, a DDPA possui um grande alcance de informação sobre desaparecidos para realização das suas pesquisas, cedida pelas Secretaria de Assistência Social, Secretaria de Saúde e todos os mecanismos da polícia fornecidos pela Secretaria de Segurança Pública do estado, como por exemplo o Disque-Denúncia[4]. Além disso, com o objetivo de alcançar uma grande parte dos cidadãos do Rio de Janeiro, a DDPA possui uma página no Facebook[5], na qual são divulgados os "cartazetes de desaparecimento". Estes cartazetes são imagens desenvolvidas pela DDPA onde são encontradas informações pessoais sobre a pessoa desaparecida, local e data do desaparecimento, e como entrar em contato com a DDPA para fornecer informações e ajudar na descoberta de paradeiro de tal pessoa (Figura 2). Após serem disponibilizados na página da delegacia, os cartazetes ficam em domínio público e podem ser compartilhados e utilizados pelos cidadãos para ajudarem na busca por informações.

Figura 2. Exemplo de cartazete disponibilizado pelo Facebook da DDPA.

Os cartazetes são desenvolvidos em editores de texto simples, sem que sejam aproveitadas automaticamente informações de um registro de ocorrência. Os membros da DDPA são responsáveis por produzir os cartazetes de desaparecidos e por divulgá-los em suas mídias sociais. Entretanto, esta prática de disponibilização em mídias sociais não fornece suporte direto para instituições parceiras como a guarda municipal, a polícia militar, o corpo de bombeiros e hospitais em geral. Não há um canal explícito e ágil de comunicação com estas instituições, com acesso rápido ao banco de informações de desaparecidos.

[3] Número do Disque-Desaparecidos: 197

[4] Número do Disque-Denúncia: (21) 2253-1177

[5] Página da DDPA no Facbook: https://www.facebook.com/DDPA-Delegacia-de-Descoberta-de-Paradeiros-972138509480582

Por exemplo, uma pessoa ao ser atropelada na rua é levada a algum hospital da cidade, porém a mesma não possui documentos que possam identificá-la e nem informações de contatos de familiares ou responsáveis. Em outra situação, um agente de saúde ou um policial, podem atender a um indivíduo sem identificação, ou desmemoriado, ou mesmo sob suspeita de rapto ou sequestro. Neste momento, o acesso à uma base de registros de desaparecidos se torna fundamental.

No entanto, no cenário atual, ao fazer o reconhecimento de uma pessoa sem identificação, as instituições parceiras devem encontrar o cartazete de informação de desaparecido divulgado pela delegacia via Facebook ou impresso, ou entrar em contato direto com a DDPA por e-mail através de uma solicitação. Recebida uma solicitação de identificação, a DDPA realizará pesquisas em diversas bases de dados (ocorrências, assistência social, sistemas prisionais etc.) a fim de identificar a pessoa solicitada. Neste sentido, a comunicação entre a DDPA e as instituições parceiras torna-se pouco eficiente, e o processo de encontrar pessoas tidas como desaparecidas, mais doloroso.

3. Tecnologia para Apoio à Busca de Desaparecidos no Rio de Janeiro

Com a necessidade de mais rapidez, agilidade e participação dos cidadãos comuns nos processos descritos anteriormente, o aplicativo móvel "Desaparecidos-RJ" (Lima, 2016) foi pensado inicialmente para ser uma ferramenta para uso por qualquer cidadão, no apoio à busca de desaparecidos, mas sobretudo os profissionais das instituições parceiras. A solução proposta se constitui de duas partes - um aplicativo mobile para a busca de informações de desaparecimento, e um sistema web, *back-end e on-line* onde os membros da DDPA registram as informações que irão alimentar o aplicativo (Figura 3).

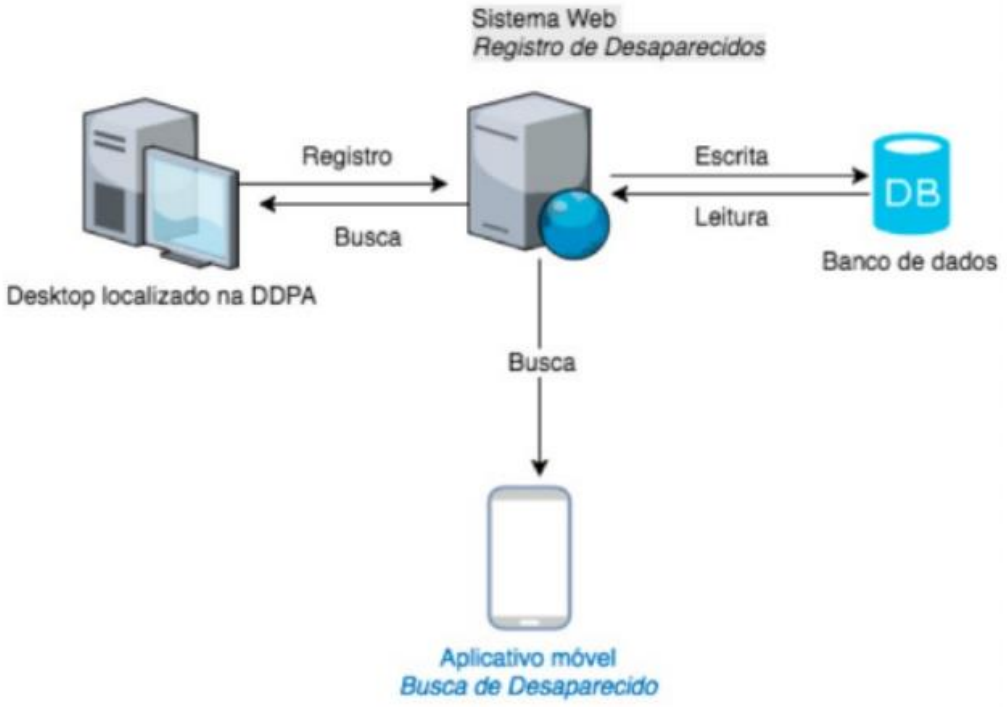


Figura 3. Arquitetura da Solução

O sistema web apresenta uma interface simples onde é possível cadastrar as informações da pessoa desaparecida como características físicas, foto, nome, detalhes (tatuagens, marcas) e outras informações. O sistema lista todos os desaparecidos cadastrados, possibilitando alterar suas informações, excluir o registro do desaparecido e gerar o cartazete, sem a necessidade de utilização de editores de textos (Figura 4). Todas as informações necessárias para a identificação de desaparecidos registrados na DPPA são armazenadas no banco de dados compartilhado do sistema web, que alimentará o aplicativo mobile.

O aplicativo mobile basicamente realiza buscas à base de dados de desaparecidos por meio de campos de informações como nome, parentes, cor de cabelo, olhos, pele, detalhes físicos etc. Ao realizar as pesquisas, o sistema irá listar informações sobre o desaparecido, possibilitando que qualquer pessoa possa identificá-lo. Ao realizar a identificação o aplicativo também permite que o usuário divulgue as informações do desaparecimento em mídias sociais, pois o sistema também é capaz de gerar cartazetes e compartilhá-los (Figura 5).

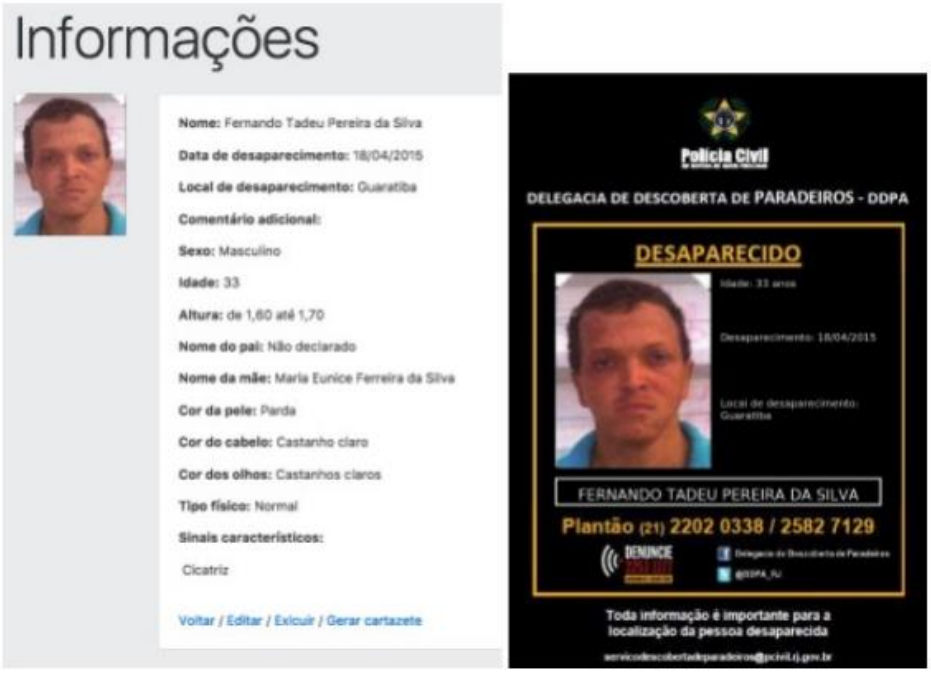


Figura 4. Cartazete Gerado Pelo Sistema Web de Registro de Desaparecidos

Figura 5. Aplicativo Desaparecidos-RJ

4. Avaliações

No intuito de averiguar a viabilidade da solução proposta, os sistemas passaram por avaliações com a DDPA e com cidadãos. A avaliação com a DDPA teve intuito de verificar se as funcionalidades implementadas nos sistemas atendiam às necessidades da delegacia quanto ao processo de busca de informações de desaparecidos e geração de cartazetes. Os testes de aceitação com agentes da DPPA foram realizados, sendo as funcionalidades aprovadas.

Para a verificação da utilidade e viabilidade do aplicativo, foi feito um teste com usuários reais, os quais deveriam responder um questionário com perguntas a respeito de funcionamento, nível de importância e usabilidade do aplicativo. Os grupos de usuários escolhidos para a realização dos testes foram alunos de graduação no curso de sistemas de informação da UNIRIO, pesquisadores e alunos de pós-graduação em informática, profissionais de saúde e assistência social. Um total de trinta (30) pessoas responderam ao questionário de perguntas e realizaram o teste do aplicativo.

O questionário de avaliação possuía cinco etapas. Na primeira etapa, foi disponibilizado o aplicativo para download e foi questionado se o download foi realizado e o aplicativo instalado com sucesso. Na sequência, foram apresentados três diferentes cenários de utilização para o aplicativo. Em cada cenário introduziu-se uma história com a necessidade de utilização do aplicativo para identificação de uma pessoa desaparecida. Após a apresentação de cada um dos cenários, os usuários foram questionados se conseguiram encontrar as informações sobre a pessoa a ser identificada e a dificuldade ou facilidade de utilizar o aplicativo. A última etapa do questionário compreende perguntas gerais sobre o uso do aplicativo incluindo sua utilidade e facilidade de uso.

Em relação aos cenários de busca de desaparecidos, a maioria dos respondentes (66,7%) conseguiu identificar a pessoa descrita no cenário. Entretanto, uma parte considerável dos usuários (33.3%) não conseguiu realizar a identificação, levando à necessidade de analisar posteriormente os comentários realizados pelos respondentes, a fim de entender as razões do insucesso da identificação.

A maioria dos comentários recebidos (35,8%) tiveram como conteúdo a atribuição de um ponto negativo à usabilidade do aplicativo em relação à busca e identificação das pessoas descritas nos cenários; 20,8% dos comentários descreveram a usabilidade do aplicativo como um ponto positivo na execução da busca; 22,6% dos comentários relataram dificuldade para buscar as pessoas dos cenários pela falta de clareza das características físicas nas buscas; 15,1% dos comentários dos usuários foram sobre não ter havido dificuldade para desempenhar as buscas; e por fim, em 5,7% dos comentários os usuários relataram que simplesmente não conseguiram realizar as buscas propostas pela cenário.

Sobre a utilidade, facilidade de uso e clareza da informação do aplicativo, a maioria dos usuários (83,3%) registaram que "o aplicativo é muito útil", em seguida, 13,3% dos usuários consideraram o aplicativo útil e 3,3% não informaram sua opinião. 56,7% dos respondentes relataram que o aplicativo foi "muito fácil de usar". A grande maioria dos usuários responderam que as informações obtidas pelo aplicativo estavam “muito claras” (63,3%).

Sobre sugestões de melhorias para o aplicativo, a maior parte (41,2%) das sugestões foram para melhorar as características físicas de busca de pessoas, ou seja, melhorar os parâmetros de busca relacionados a características físicas no formulário do aplicativo. 41,2% sugeriram melhorar a usabilidade do aplicativo, como por exemplo: incluir uma opção para selecionar diferentes características em um mesmo parâmetro de busca, disponibilizar a opção de compartilhamento de cartazete na página de detalhamento do desaparecido, e limpar a tela de preenchimento ao voltar para o formulário de uma busca. Por último, houve uma parcela (17,6%) das sugestões sobre adição de novas funcionalidades, como por exemplo, conectar o banco de dados que o aplicativo consome diretamente com o banco de dados de hospitais ou utilizar a localização do usuário para melhorar os parâmetros de busca.

5. Conclusões

O principal objetivo deste trabalho foi o desenvolvimento de uma solução tecnológica que se espera ser capaz de apoiar os processos de busca de desaparecidos executados pela Polícia Civil do Estado do Rio de Janeiro. Nesse contexto, além de agilizar e automatizar os processos citados, espera-se que a solução desenvolvida aproxime a população para colaborar com a identificação de pessoas desaparecidas, colocando em prática os conceitos da implementação da democracia digital.

Foram realizadas avaliações com a DPPA, que considerou que tanto o sistema web, quanto o aplicativo, contemplam as funcionalidades solicitadas. Os possíveis usuários do aplicativo, reportaram, em sua maioria, sua grande utilidade e a facilidade em utilizar o aplicativo para realizar a busca por pessoas desaparecidas, embora haja ainda melhorias apontadas.

Quanto à limitação da solução desenvolvida, observa-se com base no teste de usuários que: a usabilidade do aplicativo móvel dificulta a plena utilização das funcionalidades desenvolvidas e os parâmetros para busca de características físicas também dificultou a identificação de pessoas no teste. Outra limitação é existência do aplicativo apenas para a plataforma Android.

Como trabalhos futuros, sugere-se a melhoria da usabilidade do aplicativo móvel, a implementação de um algoritmo de reconhecimento facial, a disponibilização do aplicativo em outras plataformas móveis.

Agradecimentos

Os autores agradecem à Delegacia de Descoberta de Paradeiros/Polícia Civil do Estado do Rio de Janeiro, em especial ao Investigador Humberto Amaro, pela disponibilidade. Produtividade em Desenvolvimento Tecnológico e Extensão Inovadora do CNPq, Brasil processo no 305060/2016-3.

Referências

Araujo, R.; Cappelli, C.; Diirr, B; Engiel, P.; Tavares, R. (2011) “Democracia Eletrônica”. In: Pimentel, M., Fuks, H. Sistemas Colaborativos. Elsevier Brasil, pp.110-121.

Classe, T. (2017) “Processo de Descoberta de Paradeiros da Delegacia de Descoberta de Paradeiros da Polícia Civil do Rio de Janeiro”. In: Documentos do Núcleo de

Pesquisa e Inovação em CiberDemocracia, nº 0001/2017. Disponível em: http://ciberdem.uniriotec.br/wordpress/wp-content/uploads/2017/01/Ciberberm0001_2017_Processo_de_Descoberta_de_Paradeiros_DDPA.pdf.

Grandin, F.; Coelho, H. (2017) "Rio tem 33 mil desaparecidos em 15 anos; Zona Oeste e Bonsucesso concentram casos". In: Portal de Notícias do G1. Disponível em: http://g1.globo.com/rio-de-janeiro/noticia/rio-tem-33-mil-desaparecidos-em-15-anos-zona-oeste-e-bonsucesso-concentram-casos.ghtml

ISP. (2017) "Dados Oficiais". Instituto de Segurança Pública do Rio de Janeiro. Disponível em: http://www.isp.rj.gov.br/dadosoficiais.asp.

Lima, V. R. (2016) "Um sistema de informação para apoio a busca de pessoas desaparecidas no Rio de Janeiro". Trabalho de Conclusão de Curso. Bacharelado em Sistemas de Informação, Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Disponível em: http://bsi.uniriotec.br/tcc/textos/201612ViniciusRodrigues.pdf

Silva, S. P. (2005) "Graus de participação democrática no uso da Internet pelos Governos das capitais brasileiras". In Opinião Pública, v. XI(2), pp. 450-468.

Vedel, T. (2006) "The idea of electronic democracy: Origins, visions and questions". In: Parliamentary Affairs, v.59(2), pp. 226-235.

Computational Support for Updating Systematic Literature Reviews

Ramon L. Régis[1], Eber A. Schmitz[1], Marcos V. F. A. Dias[1], Priscila M. V. Lima[1]

[1]Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, RJ - Brasil

{ramon.leoncio@gmail.com, eber@nce.ufrj.br, mvfad@ufrj.br
priscila.lima@nce.ufrj.br}

Abstract. *Systematic Literature Reviews (SLR) have often been carried out but rarely updated. It is believed that the limitation of time and financial budget in research projects is one of the reasons for the lack of SLR´s updating. This study proposes an automated method of finding new documents, recommending the order in which they should be read, and suggesting new search terms to the reviewer in order to reduce the time spent in the updating process. The method proved to be effective when prioritizing documents for the search, allowing the reviewer to read more relevant than irrelevant documents, ensuring that there is less time wasted reading documents that are not useful for updating the SLR.*

1. Introduction

Systematic Literature Reviews (SLR) is a method used for summarizing a large amount of data collected and utilized in scientific papers, thesis, and books. In order to avoid biases in its systematic methodology, SLR turns the usual literature review transparent and reliable, since every method and rule used during study selection and data analysis must be explicitly described in a protocol [Kitchenham and Charters 2007].

The lifetime of an SLR is an actual issue addressed by some researchers, including the ones from the medical area [Shojania *et al.* 2007]. The process of updating an SLR is similar to the method used for conducting an SLR and, possibly, takes as much time as the first review depending on how much the original review is outdated.

The possible cause for the gap between the number of SLRs produced and that of SLRs updated is the lack of knowledge about when an SLR should be updated, a problem which is compounded by lack of resources for the updating. The main purpose of this paper is to describe an automated approach for supporting the SLR update process, with the goal of reducing the amount of human effort involved and, therefore, saving time and money in the research budget.

2. Background

2.1. Systematic Reviews

The literature review is a process executed every time a greater knowledge of some topic is needed and are a necessary component of all undergraduate monographs, master and doctoral thesis.

Literature reviews have some weakness that may bias the conclusions whenever explicit information about threats of validity is not clearly defined in the research method. Also, a literature review should be systematic, avoiding bias and exposing the methods utilized behind the review [Fink 2013]. Thus, through a systematic approach, other readers can better understand the review methods utilized by the authors, allowing them to reproduce and check the research results.

Usually, an SLR is conducted based on a protocol, defined before the reviewing. This protocol is distributed to all researchers involved to ensure that all of them will utilize the same methods for including and excluding studies and for collecting, analyzing and summarizing the data found [Biolchini *et al*. 2005].

2.2. Systematic Review Update

As the time after the SLR goes by, new problems are raised and new papers are written to address these problems. This will make an SLR old and useless sometime after its production. In the medical area, the difference between the obtained results from an SLR and its update has a significant impact on medical decision making, making clear the fact that an SLR may perish over time [Shojania *et al*. 2007].

The main goal of updating an SLR is to put effort into finding new evidence on the main research topic, whether including new terms and possibly improving the methodology. But, even that the SLR update does not locate any new evidence; it still must be considered an updated SLR [Gray and Pearce-Smith 2006].

2.3 Term-frequency vs. Inverse document frequency

Before recommending which papers should read first, the available papers must be categorized, in order that the systems may recommend the documents according to the user preferences. A common way of categorizing files is by identifying its keywords.

The keyword identification may be done through a statistical calculation named 'Term Frequency vs. Inverse Document Frequency" (*tf-idf*). The *tf-idf* calculation is made in three steps and returns a score for each word in a document representing its relevance considering the other documents in the collection [Rajaraman e Ullman 2011].

The first step calculates the term frequency *(tf)* for each word *(w)* in a given document *(d)*, dividing it by the word count *(w')* in *d*. The second step calculates the inverse document frequency *(idf)*. The *(idf)* value is obtained from the logarithm of total available documents total, divided by the documents total which belong to the set of available documents collection *(D)* which has the word *(w)* inside its body. Third, the product of *(tf)* and *(idf)* is calculated as:

$$tf - idf = tf(w, d) * IDF(w, D)$$

After these calculations, each present term in the document will receive a value, which classifies its relevance in the documents collection.

2.4. Sørensen-dice coefficient

The main action that a recommendation system must do is to associate the available content to what the user is looking for. For making this combination and for prioritizing to the user his interests, a similarity calculation, thus measuring and recommending the most similar content according to the user's profile. A method for identifying the similarity is the Sørensen-dice coefficient [Dice 1945].

The definition of the Sørensen-dice coefficient *(sc)* between two sets A and B is:

$$sc = \frac{2|A \cap B|}{|A| + |B|}$$

The value of *sc* may vary between *0* (expressing that there is no similarity between the sets *A* and *B* and *1* (expressing that the sets *A* and *B* are equivalent).

3. Approach

The proposed approach consists in reducing the time spent updating SLRs through three main steps. First **identifying** new papers that should be analyzed by the reviewers, and then **prioritizing** the most important documents for reading and finally **providing new terms** for search.

3.1 Identifying New Papers

The SLR automatic update works through the attempt of finding new evidence or verify the non-existence of new studies on the topic of research. It actually requires effort and time of the researcher for finding new evidence [Gray and Pearce-Smith 2006]. It may also be a challenge for who is doing the review because choosing the right words for mounting the search string is not a simple task and may require time and rework if the string does not cover the entire research topic [Lavallée, Robillard and Mirsalari 2014].

The digital scientific libraries search is usually done through *search string*, but the snowball technique may offer advantages over the traditional search technique, reducing the number of irrelevant studies found during the update. [Felizardo *et al* 2016]

It is possible to find documents through snowball technique that refer to the previously analyzed documents. The found documents are possibly scientific advances or just belong to the same subject, including allowing snowball to be used to conduct a new SLR [Jalali and Wohlin 2012].

Once the time invested on new documents identification and retrieval is an evident problem, the snowball is part of the presented approach for updating SLRs for its potential of reducing the effort spent by the researcher.

3.2 Prioritizing Documents for Reading

After identifying the new documents, it may happen that not all the new documents found to belong to the SLR main theme. Thus they're not relevant to the SLR update, influencing in a greater time invested by the reviewer while trying to classify what studies must be included or excluded from the review.

Even after the title, abstract and keywords analysis, some documents might need to have its whole body analyzed to allow the researcher to have a final judgment on whether the document will belong or not to the SLR update, requiring more time for the full read.

This SRL update approach proposes a recommender system for helping the researcher on reading first what is interesting for its research topic, reducing the time spent on irrelevant studies and ensuring that the important documents are being prioritized due to time or budget limitations for the research [Cohen, Ambert and McDonagh 2009].

3.3 New Search String

As new theories and concepts is rising, and as the science advances, new terms rise too. These terms must be included in the review to ensure that the whole topic is being covered [Gray and Pearce-Smith 2006].

For each search that is done, if an important term is found, it must be included and adjusted to the search string. Keywords are not immediately obvious but a great drawback in an SLR is to find new keywords after all the reviewing was done, implying rework [Lavallée, Robillard and Mirsalari 2014].

The final step of the proposed approach is to identify new keywords for search string inclusion, based on the user implicit preferences, through an analysis of the terms present on each important document, preventing the rework and increasing the SLR validity.

3.4 Process of the Proposed Method

The proposed method involves activities that can be automated, facilitating the conduct of the SLR update. The activities that make up the whole process behind the proposed method were modeled and are represented in Figure 1.

The first step for classifying what paper the researcher should read first is to grade the quality of each document in the initial revision. In this scenario, we consider quality as a measure that defines how much the subject of the paper is aligned with the review main theme. This grade is a number that may vary from *1* to *10*, where *1* represents that the paper has a minimum importance for the SLR and 10 means that the paper is fully addressed to the main topic being reviewed. A grade containing 0 as its value is not being considered once a paper with grade 0 represents a paper that should not be in the initial review.

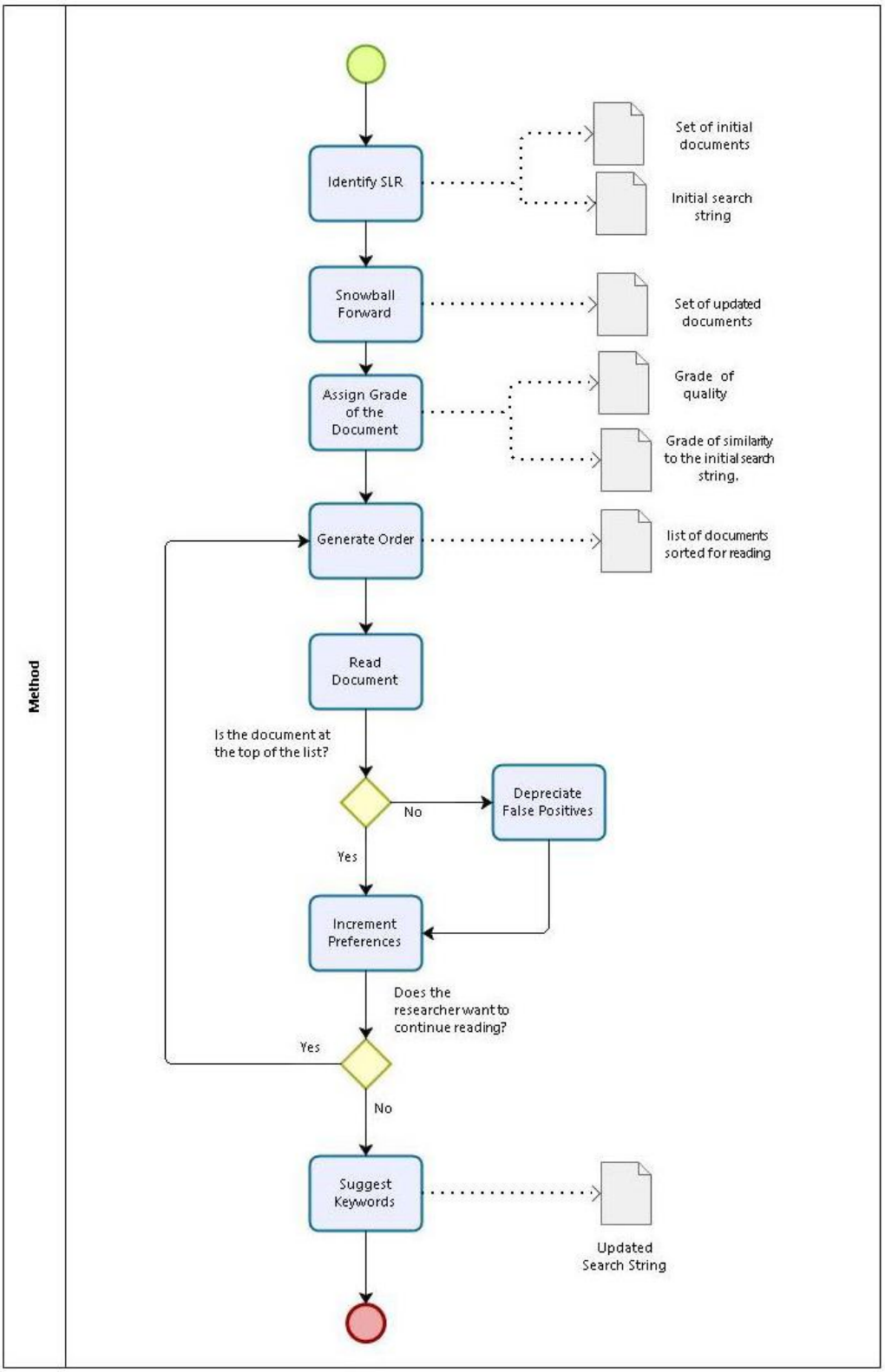


Figure 1: Overview of the proposed approach

Once the review original documents grades are known, it is possible to get their 'children' papers, i.e. the papers that cite them. After this, each children paper receives its father's grade. For each found document, its keywords must be classified according to their relevance among the all the documents. We used only the top *100* keywords.

The next step is to calculate the similarity between the found document and the initial *search string*. This is done through the comparison between the initial *search string* words and the previous document classified words.

The last step of providing the initial order is to calculate the final score. Once the final score is calculated, the reading order is presented to the reader.

After a document is read, the user preferences must be changed. So the recommender system may reorder its documents list to provide the new reading order which the reviewer should follow.

4. Experiment

In order to evaluate the proposed method, we executed an experimental study. The experiment consisted of the application of the proposed method to an existent SLR. The SLR selection criteria were: (1) an SLR of at least two years of publication, (2) available for download and (3) the SLR main subject was of knowledge to the experiment driver.

The selected SLR, [Rafi *et al*. 2012], contained 24 documents. The snowball forward of these documents returned a total of 900 documents. From these *900* documents, *63* were duplicates and *134* could not be downloaded for financial, license or availability reasons. From the remaining documents, *279* were published before *2011*, the SLR updating. The initial SLR had another rule as inclusion criteria: only studies in English should be included. Thus *57* documents were removed containing non-Latin characters or which were written in Portuguese, Spanish, French or German.

The remaining *360* documents were ordered. After the initial order was given, a sample of *100* documents was selected. The whole sample of documents was read so that they could be classified as being useful or not to the SLR update.

For the sample selection criteria, the first *20* documents present at the initial given order were read. Following the approach, these *20* first ones should be considered useful documents for the SLR update. Also, the last *20* documents from the initial list were chosen, which following the approach tend to be documents which are useless to the SLR update. Randomly *60* documents were selected from the middle of the initial order list, composing the sample of 100 documents. From the *100* selected documents, *31* were considered useful to the SLR update and *69* were considered irrelevant to the updating.

One hundred iterations were made to the ordered list, selecting each document on the recommended order such that a relation between the amount of useful and irrelevant documents could be identified over time. For each iteration, the documents ordered list was recorded, such that the documents' different positions could be identified over the iterations, conforms Figure 2.

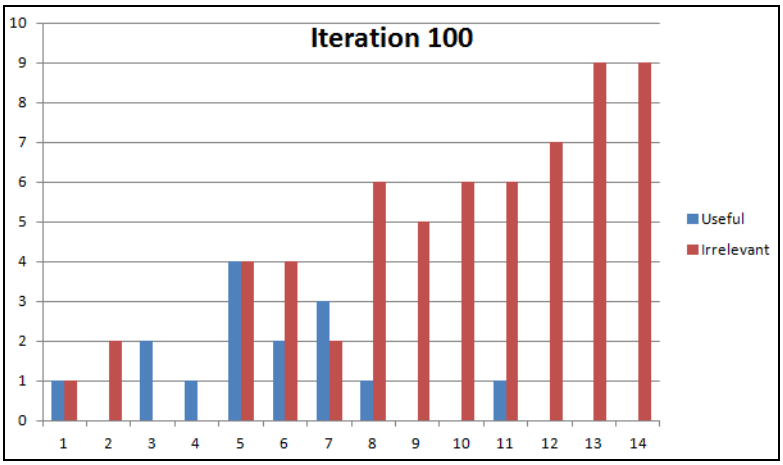


Figure 2: Status after Iteration 100

5. Considerations

The proposed approach showed to improve the quality of the review conduction, based on the objective of helping the reviewer to read the relevant documents first.

The documents recommendation has a greater effect during the first iterations rather than the last iterations due to the less relevant (but still useful) papers having similar keywords to the documents considered irrelevant.

The words suggested during the iterations proved to be relevant words to the user. It is not indispensable, however, that the given words must be included in a *search string* but, they must be analyzed in order to identify if the new words have some relevance for finding new documents.

Finally, the effect of the recommender system is greater at the beginning and decreases its effectiveness through the iterations. It ensures that the reviewer is going to read relevant papers first, before getting dragged down reading irrelevant papers.

5.1. Threats to Validity

The proposed method showed to be a great tool for helping reviewers to read what is important first, during an SLR update. But due to the fact that conducting the experiment includes the selection of an out of date SLR, the SLR main topic knowledge, the analysis of hundreds of scientific papers and the full reading of part of them, it would take more time than the time available for this research to repeat the experiment and confirm the results we obtained.

5.2. Future Work

The preferences collected explicitly could improve the effectiveness of the method if the reviewer could choose some specific keywords as his preference before the initial order. If the reviewer, that usually acts into a specific topic of study, keeps his preferences to be used for future recommendations, these preferences collected implicitly could improve the effectiveness of the method. Finally, a suggestion for new keywords could use a dictionary or maybe some natural language processing technique to improve the accuracy on showing relevant words to the user.

References

Biolchini, J. et al. Systematic Review in Software Engineering. **Engineering**, v. 679, n. May, p. 1–31, 2005.

Cohen, A. M.; Ambert, K.; McDonagh, M. Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update. **Journal of the American Medical Informatics Association**, v. 16, n. 5, p. 690–704, 2009.

Dilce, L. R. Measures of the Amount of Ecologic Association Between Species Author (s): Lee R . Dice Published by : Wiley Stable URL : http://www.jstor.org/stable/1932409 Accessed : 08-04-2016 13 : 33 UTC Your use of the JSTOR archive indicates your acceptance of th. **Ecology**, v. 26, n. 3, p. 297–302, 1945.

Felizardo, K. R.; Mendes, E.; Kalinowski, M.; Souza, É. F.; Vijaykumar, N. L. Using forward snowballing to update systematic reviews in software engineering. In: **International Symposium On Empirical Software Engineering and Measurement**, 2016, Ciudad Real, Espanha. Proceedings... 2016.

Fink, A. **Conducting Research Literature Reviews: From the Internet to Paper**. [s.l: s.n.].

Kitchenham, B.; Charters, S. Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3. **Engineering**, v. 45, n. 4ve, p. 1051, 2007.

Gray, J M.; Pearce-Smith, N. Systematic reviews: when is an update an update? **Lancet**, v. 367, n. 9528, p. 2058, 2006.

Jalali, S.; Wohlin, C. Systematic Literature Studies: Database Searches vs. Backward Snowballing. **International Symposium on Empirical Software Engineering and Measurement**, p. 29–38, 2012.

Lavallée, E.; Robillard, P. N.; Mirsalari, R. Performing systematic literature reviews with novices: An iterative approach. **IEEE Transactions on Education**, v. 57, n. 3, p. 175–181, 2014.

Rafi, D. M. et al. Benefits and Limitations of Automated Software Testing : Systematic Literature Review and Practitioner Survey. **Automation of Software Test (AST), 2012 7th International Workshop on**, p. 36–42, 2012.

Rajaraman, A.; Ullman, J. D. Data Mining. **Mining of Massive Datasets**, v. 18 Suppl, p. 114–142, 2011.

Shojania, K. G. et al. Updating Systematic Reviews--Technical Review 16. **Agency for Healthcare Research and Quality**, v. 7, n. 16, p. 135, 2007.

Análise e Integração dos Dados Abertos do Sistemas de Transporte Público de Curitiba

Elis Cassiana Nakonetchnei[1], **Nádia Puchalski Kozievitch**[1]
Anelise M. Fonseca[1], **Ana C. K. Vendramin**[1], **Keiko V. O. Fonseca**[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba – PR – Brasil

elisan@alunos.utfpr.edu.br, nadiap@utfpr.edu.br, anelise@utfpr.edu.br

cristina@dainf.ct.utfpr.edu.br, keiko@utfpr.edu.br

Abstract. *The increase in population generated a demand for new solutions in order to manage the urban centers. In parallel, the advance of technology has made it easier to obtain data about the systems which make up a city. The result of this scenario was the creation of a large amount of data, that is growing every day, and needs a pre-process to be able to transform them into integrated information. This article aims to analyze the public transportation data of Curitiba, made available by official sources, in order to find a way to optimize the integration of new data obtained with the existing data in the base of the EUBra-BIGSEA project, listing in parallel the challenges encountered during the process.*

Resumo. *O aumento da população gerou uma demanda por novas soluções para administrar os centros urbanos. Em paralelo, o avanço da tecnologia proporcionou uma maior facilidade para se obter dados sobre os sistemas que compõe uma cidade. O resultado deste cenário foi a criação de um grande volume de dados, que cresce cada dia mais, sendo necessário pré-processá-lo para ser possível transformá-los em informação integrada. Este artigo tem como objetivo analisar dados sobre o transporte público de Curitiba, disponibilizados por fontes oficiais, buscando uma forma de otimizar a integração dos novos dados obtidos aos dados existentes na base do projeto EUBra-BIGSEA, listando paralelamente os desafios encontrados durante o processo.*

1. Introdução

O constante crescimento dos centros urbanos traz consigo diversos empecilhos que precisam ser superados para fornecer o máximo conforto possível para seus habitantes. Sendo assim, é necessário criar novas maneiras de administrar uma cidade, conciliando o funcionamento de diversos sistemas de forma que seu desempenho seja capaz de atender melhor a todos. Nesse cenário, surgiu o conceito de *smart city* (ou cidades inteligentes), que embora não possua uma definição única, pode ser entendida como a integração da tecnologia numa abordagem estratégica de sustentabilidade, bem estar dos cidadãos e desenvolvimento econômico [Kozievitch et al. 2016].

Entre as preocupações existentes nas cidades inteligentes, o sistema de transporte público é um destaque e, segundo [Kozievitch et al. 2016], pode ser considerado um fator crítico para o funcionamento de uma cidade. A eficiência do seu desempenho auxilia

na diminuição da frota e também provê provê uma integração social, pois para algumas pessoas o sistema de transporte público é o único meio de deslocamento entre suas tarefas diárias [Weigang et al. 2001]. Porém, em algumas regiões a falta de investimento no transporte coletivo e a atual facilidade em obter crédito para a aquisição de veículos tornam o transporte individual mais atrativo [Secron et al. 2016].

Em Curitiba [1], o crescimento sustentável da cidade é uma preocupação antiga, que remonta ao século XIX [2]. No caso do transporte, existe a priorização do transporte coletivo sobre o individual, fazendo com que o sistema funcione em um conceito de rede integrada. A rede integrada que permite ao usuário o deslocamento entre pontos da cidade, utilizando mais de um tipo de ônibus [Sequinel et al. 2002].

Pensando no crescimento da cidade, a Prefeitura de Curitiba possui uma iniciativa de dados abertos [3], onde é possível encontrar dados relacionados a diversas áreas da cidade. A fonte destes dados são órgãos oficiais, como por exemplo o Instituto de Pesquisa Planejamento Urbano de Curitiba (IPPUC) [4] e a Urbanização de Curitiba [5] (URBS), responsável pelo transporte público na cidade. Estas iniciativas tem como objetivo disponibilizar aos cidadãos dados sobre a gestão municipal para livre utilização pela sociedade, podendo ser reutilizados para produzir novas informações e aplicações digitais.

No entanto, mesmo existindo uma quantidade significativa de dados disponíveis à população, eles se encontram separados em bases de dados distintas, havendo pouca ligação direta entre as áreas. Segundo [de Sá Rodrigues et al. 2016], os motivos para não integração de dados são: (1) a grande quantidade de conteúdo divulgado, pois gerenciar um grande volume de dados exige muito trabalho, sendo considerado um dos grandes desafios da computação [6]; e (2) a falta de um vocabulário comum entre as fontes desses dados, necessitando de uma fase de adaptação dos dados. Esta segmentação dos dados dificulta a sua manipulação quando existe a necessidade de mesclar um ou mais conjuntos, reduzindo o potencial de aplicação dos mesmos em busca de melhorias para cidade.

Considerando o problema descrito anteriormente, [Barczyszyn 2015] idealizou e iniciou a construção de uma base que integra dados de diversas áreas de Curitiba, facilitando a análise e a conversão em conhecimento, o que permite entender o comportamento da cidade de forma dinâmica. Esta base foi adotada pelo projeto EuBra-BIGSEA [7], uma colaboração entre o Brasil e a Europa de pesquisa cientifica sobre *Big Data* através de aplicações centradas na nuvem, permitindo o processamento massivo de dados para resolver problemas de descoberta de conhecimento.

Por fim, o trabalho descrito neste artigo tem como um dos principais objetivos buscar uma forma de otimizar a integração de novos dados sobre o sistema de transporte público, com os dados já existentes na base do projeto EUBra-BIGSEA, listando

[1] http://www.curitiba.pr.gov.br/ (acessado em 21/08/2016)

[2] http://www.curitiba.pr.gov.br/conteudo/historia/1615 (acessado em 13/09/2017)

[3] http://www.curitiba.pr.gov.br/dadosabertos/ (acessado em 17/11/2016)

[4] http://www.ippuc.org.br (acessado em 17/10/2016)

[5] https://www.urbs.curitiba.pr.gov.br/ (acessado em 17/10/2016)

[6] http://www.sbc.org.br/documentos-da-sbc/send/141-grandes-desafios/802-grandesdesafiosdacomputaonobrasil (acessado em 29/08/2016)

[7] http://www.eubra-bigsea.eu/ (acessado em 29/08/2016)

os desafios encontrados durante o processo de integração dos dados para auxiliar futuras integrações.

Este documento está organizado da seguinte maneira: a Seção 2 apresenta uma comparação entre os novos dados obtidos e os contidos na base de dados utilizada, a Seção 3 lista os principais desafios encontrados e a Seção 4 apresenta a conclusão e trabalhos futuros.

2. Comparação entre os Conjuntos de Dados

Apesar de ambas as bases (EuBra-BIGSEA e URBS) representarem o sistema de transporte público de Curitiba, a organização dos dados dos dois conjuntos (IPPUC - EUBra-BIGSEA e URBS) não possui o mesmo padrão. Para os dados provenientes da URBS, existem quatro tabelas para representar suas informações, ilustradas na Figura 1. No projeto EUBra-BIGSEA, existem três tabelas, guardadas em PostGIS (ilustradas na Figura 2) para representar as informações sobre o transporte.

Ao comparar quais dados são utilizados para representar uma mesma informação, é possível notar que no caso do projeto EuBra-BIGSEA, existe uma diferença entre pontos de ônibus e terminais, enquanto nos dados vindos da URBS, não existe uma tabela individual para terminais. Em contrapartida, a URBS possui uma tabela para retratar os trajetos realizados pelos ônibus, que em conjunto com a tabela de linhas de ônibus, retratam a mesma informação representada pela tabela linha de ônibus do projeto EuBra-BIGSEA. Por último, é possível notar que existe uma tabela resultante dos dados da URBS, chamada veículos, que retrata a geolocalização de cada ônibus em circulação, com data e hora, que não existe no projeto EuBra-BIGSEA.

(A)

```
CREATE TABLE urbs_pontos
(
    cd_linha character varying(25) NOT NULL,
    nome character varying(100),
    num_ponto character varying(25) NOT NULL,
    sequencia character varying(25) NOT NULL,
    grupo character varying(25),
    sentido character varying(100) NOT NULL,
    tipo character varying(100),
    latitude double precision,
    longitude double precision,
    gid serial NOT NULL,
    geom geometry(Point,4326),
    CONSTRAINT pontos_pkey PRIMARY KEY (cd_linha, num_ponto, sequencia, sentido)
)
```

(B)

```
CREATE TABLE urbs_veiculos
(
    prefixo character varying(25) NOT NULL,
    hora time without time zone NOT NULL,
    cd_linha character varying(25),
    adap_cadeirante character varying(25),
    latitude double precision NOT NULL,
    longitude double precision NOT NULL,
    gid serial NOT NULL,
    geom geometry(Point,4326),
    CONSTRAINT veiculos_pkey PRIMARY KEY (prefixo, hora, latitude, longitude)
)
```

(C)

```
CREATE TABLE urbs_path
(
    cd_linha character varying(25) NOT NULL,
    lat double precision,
    lon double precision,
    gid serial NOT NULL,
    geom geometry(Point,4326)
)
```

(D)

```
CREATE TABLE urbs_linhas
(
    cd_linha character varying(25),
    linha character varying(100),
    somente_cartao character(1),
    categoria character varying(100),
    gid serial NOT NULL,
    geom geometry(LineString,4326),
    CONSTRAINT giovane_urbs_linhas_pkey PRIMARY KEY (gid)
)
```

Figura 1. Tabelas com dados da URBS.

(A)

```
CREATE TABLE transporte.terminal_de_transporte
(
  gid serial NOT NULL,
  cd_equi numeric,
  tema character varying(50),
  id_equip numeric(10,0),
  equipament character varying(150),
  tipo_equi character varying(60),
  dep_admin character varying(60),
  pre_nome character varying(150),
  nome character varying(150),
  sigla_equi character varying(50),
  conveniado character varying(3),
  nome_abrev character varying(60),
  nome_mapa character varying(111),
  cd_rua character varying(5),
  nome_rua character varying(106),
  nome_ruano character varying(150),
  num_pred character varying(15),
  compl_end character varying(254),
  indfiscal character varying(20),
  cd_bairro numeric(10,0),
  bairro character varying(30),
  quadr_equi character varying(4),
  cd_regiona numeric(10,0),
  regional character varying(16),
  func_manha character varying(3),
  func_tarde character varying(3),
  func_noite character varying(3),
  func_24hr character varying(3),
  telefone character varying(15),
  ramal character varying(15),
  email character varying(60),
  site character varying(150),
  dt_inaugur date,
  desativado character varying(1),
  observacao character varying(254),
  fonte character varying(250),
  dt_atualiz date,
  coord_e numeric,
  coord_n numeric,
  geom geometry(Point,4326),
  CONSTRAINT terminal_de_transporte_pkey PRIMARY KEY (gid)
)
```

(B)

```
CREATE TABLE transporte.linha_de_onibus
(
  gid serial NOT NULL,
  objectid numeric(10,0),
  layer character varying(254),
  cd_categor character varying(25),
  categoria character varying(100),
  cd_linha character varying(25),
  linha character varying(100),
  data character varying(25),
  fonte character varying(50),
  seta_senti character varying(50),
  shape_len numeric,
  geom geometry(MultiLineString,4326),
  CONSTRAINT linha_de_onibus_pkey PRIMARY KEY (gid)
)
```

(C)

```
CREATE TABLE transporte.ponto_de_onibus
(
  gid integer NOT NULL DEFAULT nextval('transporte.ponto_de_onibus_gid_seq1'::regclass),
  entity character varying(16),
  handle character varying(16),
  layer character varying(254),
  lyrfrzn smallint,
  lyrlock smallint,
  lyron smallint,
  lyrvpfrzn smallint,
  lyrhandle character varying(16),
  color smallint,
  entcolor smallint,
  lyrcolor smallint,
  blkcolor smallint,
  linetype character varying(254),
  entlinetyp character varying(254),
  lyrlntype character varying(254),
  blklinetyp character varying(254),
  elevation numeric,
  thickness numeric,
  linewt smallint,
  entlinewt smallint,
  lyrlinewt smallint,
  blklinewt smallint,
  refname character varying(254),
  ltscale numeric,
  extx numeric,
  exty numeric,
  extz numeric,
  docname character varying(254),
  docpath character varying(254),
  doctype character varying(32),
  docver character varying(16),
  geom geometry(PointZM,4326),
  CONSTRAINT ponto_de_onibus_pkey PRIMARY KEY (gid)
)
```

Figura 2. Tabelas do projeto EUBra-BIGSEA.

Uma outra diferença nos dados que pode ser citada é na descrição da categoria de ônibus, como ilustra a Figura 3: o conjuntos de dados presentes é distinto.

	urbs character varying(100)
1	ALIMENTADOR
2	CIRCULAR CENTRO
3	CONVENCIONAL
4	ENSINO ESPECIAL
5	EXPRESSO
6	INTERBAIRROS
7	JARDINEIRA
8	LIGEIRÃO
9	LINHA DIRETA
10	MADRUGUEIRO
11	TRONCAL

	eubrabigsea character varying(100)
1	ALIMENTADOR
2	CIRCULAR CENTRO
3	CONVENCIONAL
4	EXPRESSO
5	EXPRESSO LIGEIRÃO
6	INTERBAIRROS
7	INTERHOSPITAIS
8	LINHA DIRETA
9	METROPOLITANO
10	TRONCAL
11	TURISMO

Figura 3. Categorias de ônibus.

Já na categoria linhas de ônibus, existe uma diferença de 83 linhas de ônibus entre as duas bases de dados, como ilustra a Tabela 1. Nas divergências encontradas entre as bases de dados, ilustradas nas Tabelas 2 e 3, foi possível encontrar dois casos:

1. Códigos de linhas de ônibus de uma base semelhantes a um código de outra base. EXEMPLO: código 207-1 (Projeto EUBra-BIGSEA) e o código 207 (URBS).
2. Códigos de linhas de ônibus de uma base que não são semelhantes a nenhum código de outra base. EXEMPLO: código 105 (Projeto EUBra-BIGSEA) e o código 27 (URBS).

Tabela 1. Quantidade de linhas de ônibus.

	Comum	Divergente	Total
URBS	266	147	413
EUBra-BIGSEA	266	230	496

Tabela 2. Linhas de ônibus divergentes (EUBra-BIGSEA).

105	606-1	921	B02	B61	C20	D66	F03	G13	J15
182-1	606-2	922	B05	B72	C22	E01	F05	G71	J16
182-2	650-1	923-1	B06	B73	C23	E02	F12	G72	J62
207-1	650-2	923-2	B11	B74	C25	E05	F13	G73	K11
216-1	650-3	A01	B13	B75	C26	E11	F14	H01	K71-1
216-2	684-1	A02	B14	B76	C27	E21	F15-1	H11	K71-2
219-1	684-2	A06	B15-1	B77	C28	E31	F15-2	H12	L11
219-2	684-3	A07	B15-2	B78	C30	E62	F15-3	H20	L71
231-1	720-1	A11	B17	B79	C41	E63	F16	H21	N70
231-2	720-2	A13	B18	B80	C63	E64	F17	I20	N71
301	805	A14	B19	B81	C66	E65	F18	I21	N72
302-1	806	A16-1	B20	B82	C72	E66	F19	I30	N73
302-2	814-1	A16-2	B22	B83	D11	E67	F21	I31	O71
304-1	814-2	A17	B23	C03	D12	E68	F22	I32	O72
304-2	815-1	A21	B24	C04-2	D13-1	E70	F24	I40	O73
304-3	815-2	A22	B26	C05	D13-2	E71	F25	I41	O74
548-1	822-1	A31	B27	C11	D14	E72	F26-1	I50	O75
548-2	822-2	A32	B28	C12	D16	E73	F26-2	I71	P63
601-1	824	A72	B29	C13	D21	E75	F27	I90	P64
601-2	902-1	A73	B31	C15	D22	E76	F72	I91	P65
603-1	902-2	A77	B33	C16	D23	E77	F73	J12	R71
603-2	915-1	A78	B34	C17	D31	E78	G11	J13	X11-1
605	915-2	B01	B35	C18	D61	F01	G12	J14	X20-2

Tabela 3. Linhas de ônibus divergentes (URBS).

27	229	304	399	565	682	706	813	989	X19
164	231	308	489	567	684	710	814	994	X20
188	234	309	494	596	686	716	815	995	X21
189	235	314	496	597	687	720	816	996	X22
194	245	319	497	598	689	731	822	997	X23
195	256	333	499	599	691	775	889	998	X24
196	259	339	509	601	692	788	893	999	X25
197	289	389	510	603	693	789	894	WWW	X26
198	294	391	514	608	694	795	895	X01	X27
199	295	392	517	609	695	796	896	X04	X30
207	296	394	519	626	696	797	897	X10	X31
209	297	395	527	634	697	798	898	X11	Z02
210	298	396	543	650	698	799	899	X13	
216	299	397	548	669	699	802	902	X15	
219	302	398	549	679	705	809	915	X17	

Já na categoria pontos de ônibus, a base de dados que apresenta maior divergência, sendo a primeira delas, os sistemas de georreferenciamento adotados: latitude/longitude pela URBS e Universal Transversa de Mercator (UTM) pelo EUBra-BIGSEA. Esta diferença foi transposta através de funções do PostGIS. Observando, com auxílio de um mapa, a distribuição dos pontos de ônibus das duas bases pela cidade, como ilustra a Figura 4, é possível notar que os dados provenientes da URBS dificilmente ultrapassam os limites de Curitiba, ao contrário do projeto EUBra-BIGSEA que possui dados sobre pontos de ônibus fora dos limites da cidade. Numericamente, a diferença entre o total de pontos de ônibus encontrados em cada base de dados é de 4564, ilustrada na Tabela 4. Porém, ao se limitar apenas aos pontos que se encontram dentro dos limites da cidade, a diferença diminui para 1809.

Tabela 4. Comparação entre a quantidade de pontos de ônibus.

Fonte	URBS	EUBra-BIGSEA	Diferença
Quantidade Total	18510	23074	4564
Quantidade em Curitiba	16590	18399	1809

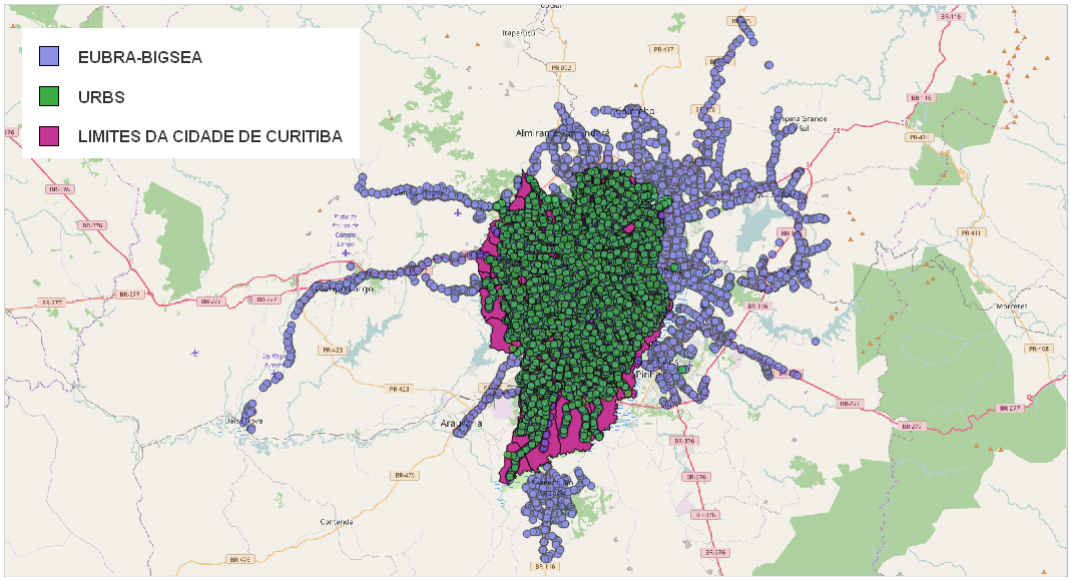


Figura 4. Distribuição dos pontos de ônibus.

3. Desafios Encontrados

O projeto EUBra-BIGSEA trabalha com dados provenientes de duas fontes ligadas à Prefeitura da Cidade de Curitiba. Porém, apesar de ambas as fontes serem oficiais, não existe uma padronização no armazenamento dos dados, ou seja, cada fonte opta pela forma que lhe é mais conveniente para armazenar as informações que possui sobre uma mesma área. Consequentemente, durante a integração, foram encontrados desafios, na forma de problemas de qualidade de dados e divergências entre as fontes, que atrasaram o andamento do projeto, pois exigiram um pré-processamento para torná-los compatíveis. Os desafios encontrados durante a integração dos dados estão listados a seguir:

1. Precisão: Nível de detalhamento que uma fonte escolhe para representar uma entidade no banco de dados. Neste tipo de problema (exemplificado na Figura 5), cada fonte escolhe atributos diferentes para representar uma mesma entidade;
2. Consistência: Inexistência de conflitos em um banco de dados específico. Neste tipo de problema (exemplificado na Figura 5), cada fonte escolhe um tipo diferente para representar um mesmo atributo existente em ambas as bases;
3. Exatidão: Exatidão que um dado armazenado possui em relação ao "valor" real que ele representa. Neste tipo de problema (exemplificado na Figura 6), a forma que a fonte armazena o valor de um atributo armazenado na fonte não é exatamente igual ao seu valor no "mundo real";
4. Diferença entre as quantidades de dados que cada fonte possui sobre um mesmo tópico (exemplificado na comparação de pontos de ônibus).

4. Conclusão

Este artigo apresentou uma comparação entre os dados de mobilidade usando a base de dados de um projeto que visa a integração de dados de diversas áreas, e novos dados oficiais sobre o transporte público da cidade de Curitiba.

Foram verificadas divergências nas bases de dados com relação à precisão, exatidão, consistência e número de tuplas. Vários destes problemas podem ser solucionados

```
CREATE TABLE urbs_linhas
(
  cd_linha character varying(25),
  linha character varying(100),
  somente_cartao character(1),
  categoria character varying(100),
  gid serial NOT NULL,
  geom geometry(LineString,4326),
  CONSTRAINT giovane_urbs_linhas_pkey PRIMARY KEY (gid)
)
```

```
CREATE TABLE transporte.linha_de_onibus
(
  gid serial NOT NULL,
  objectid numeric(10,0),
  layer character varying(254),
  cd_categor character varying(25),
  categoria character varying(100),
  cd_linha character varying(25),
  linha character varying(100),
  data character varying(25),
  fonte character varying(50),
  seta_senti character varying(50),
  shape_len numeric,
  geom geometry(MultiLineString,4326),
  CONSTRAINT linha_de_onibus_pkey PRIMARY KEY (gid)
)
```

Figura 5. Exemplo de Problema do Tipo de Precisão e Consistência.

codigo_linha character varying(25)	nome_urbs character varying(100)	nome_bigsea character varying(100)
180	ÁGUA VERDE/ ABRANCHES	ÁGUA VERDE-ABRANCHES

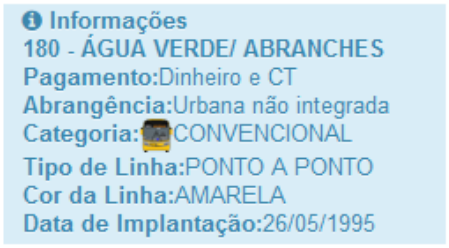


Figura 6. Exemplo de Problema do Tipo de Exatidão.

com funções geométricas de bancos de dados espaciais, porém, é necessário verificar possíveis registros reduntantes para solucioná-las.

Como trabalhos futuros podemos citar o processo de adequação dos novos dados obtidos para integrá-los com os dados já existentes na base projeto, a comparação com outras fontes, o uso dos dados para resolver problemas da cidade, e a sugestão de padrões para a municipalidade.

Agradecimentos

Agradecimentos à Prefeitura de Curitiba, URBS, Instituto de Pesquisa e Planejamento Urbano de Curitiba e ao *EU-BR EUBra-BigSea project (MCTI/RNP 3rd Coordinated Call)* pelo fornecimento dos dados.

Referências

[Barczyszyn 2015] Barczyszyn, G. L. (2015). *Integração de dados geográficos para planejamento urbano da cidade de Curitiba*. Trabalho de Conclusão de Curso, Universidade Tecnológica Federal do Paraná.

[de Sá Rodrigues et al. 2016] de Sá Rodrigues, G., Paixão, G., and Brito, A. (2016). Uma aplicação interligando dados de gps com linked geo data. *III Anais da Escola Regional de Sistemas de Informação do Rio de Janeiro*, pages 83–88.

[Kozievitch et al. 2016] Kozievitch, N. P., Gadda, T. M. C., Fonseca, K. V. O., Rosa, M. O., Gomes-Jr, L. C., and Akbar, M. (2016). Exploratory analysis of public transportation data in Curitiba. *XXXVI Anais do Congresso da Sociedade Brasileira de Computação*, pages 1656–1667.

[Secron et al. 2016] Secron, T. M., da Silva, E. R., de Farias, C. M., and Cruz, T. (2016). Sigaciente: Uma ferramenta para inferência do trânsito e de rotas seguras baseada em dados sociais. *III Anais da Escola Regional de Sistemas de Informação do Rio de Janeiro*, pages 58–65.

[Sequinel et al. 2002] Sequinel, M. C. M. et al. (2002). O modelo de sustentabilidade urbana de curitiba: um estudo de caso.

[Weigang et al. 2001] Weigang, L., Yamashita, Y., da Silva, O. Q., XiJun, D., dos Prazeres, M. Â. T., and de Oliveira, D. C. S. (2001). Implementação do sistema de mapeamento de uma linha de ônibus para um sistema de transporte inteligente. In *Anais do XXI Congresso da Sociedade Brasileira de Computação, Seminário Integrado de Software e Hardware (SEMISH)*, pages 72–85.

Valor Presente Líquido em Projetos de *Software* com e sem restrição de Recursos

Isac Mendes Lacerda[1] e Eber Assis Schmitz[1]

[1]Programa de Pós-Graduação em Informática (PPGI)
Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ - Brasil

isac.mendes@gmail.com, eber@nce.ufrj.br

Abstract. *This paper compares two different approaches to the problem of maximizing the objective function Net Present Value (NPV) in a hypothetical software project. This function is evaluated under the Critical Path Method (CPM) approach, without restriction of resources, and also under the resource-constrained approach in [Denne and Cleland-Huang 2003]. The generation of the data counted on the use of algorithms elaborated in the language R, in order to dynamize the simulations and analyses. With this, it was possible to verify the best results between the two approaches studied.*

Resumo. *Este trabalho compara duas diferentes abordagens para o problema de maximização da função objetivo Valor Presente Líquido (VPL) em um projeto hipotético de software. Essa função é avaliada sob a abordagem do Critical Path Method (CPM), sem restrição de recursos, e também sob a abordagem tratada em [Denne e Cleland-Huang 2003], com restrição de recursos. A geração dos dados contou com a utilização de algoritmos elaborados na linguagem R, de modo a dinamizar as simulações e análises. Com isso, foi possível verificar os melhores resultados entre as duas abordagens estudadas.*

1. Introdução

Engenharia de *software* pode ser definida como a aplicação de métodos científicos ou empíricos com finalidade de produzir *software* de qualidade dentro de prazo e custos previsíveis [Pressman 2006]. Nesse sentido, entende-se que a engenharia de *software* é uma grande caixa que compreende todos os esforços de construção, tanto os de natureza técnica como os de natureza de gestão. Um exemplo disso é a classificação que o *Unified Process* apresenta para suas nove disciplinas: disciplinas de engenharia, mais voltadas ao produto; e disciplinas de suporte, mais voltadas ao apoio e à gestão do projeto [Jacobson *et al.* 1999]. Naturalmente, essa é uma visão moderna da engenharia de *software* que sofreu adaptações ao longo do tempo. Tomando uma definição de engenharia de *software* dos anos 60, é possível notar que algumas preocupações modernas são deixadas de lado, como por exemplo, pontualidade, satisfação do cliente, unidades de medida e processo amadurecido. A definição dada por Fritz Bauer em 1969 [Naur e Randall 1969] considera que a engenharia de *software* é a criação e a utilização de sólidos princípios de engenharia a fim de obter *softwares* econômicos que sejam confiáveis e que trabalhem eficientemente em máquinas.

Em 1993, o *Institute of Electrical and Eletronics Engineers* (IEEE) apresentou uma definição mais ampla, destacando outras legítimas preocupações da engenharia de *software*. Segundo o [IEEE 1993], a atividade inclui também a aplicação de uma abordagem sistemática, disciplinada e quantificável, para o desenvolvimento, operação e manutenção do *software*. Diante disso, entende-se que a engenharia de *software* inclui aspectos econômicos e financeiros

para um projeto, especialmente sobre a decisão de fazê-lo ou não. Para o tratamento desses aspectos, existem vários métodos de apoio, comuns a projetos de qualquer área. Entre eles, aqui também chamados de funções objetivo, estão o Valor Presente Líquido (VPL), o Autofinanciamento e o Ponto de Equilíbrio [Denne e Cleland-Huang 2003].

Tais funções objetivo podem ser usadas não só para a seleção de diversos projetos (visão de portfólio), mas também para a escolha de quais módulos dentro de um único projeto devem ser feitos e em que ordem. Esse destaque é importante, pois a visão moderna de engenharia de *software* também inclui como uma de suas principais práticas o desenvolvimento iterativo*,* que tem por compromisso fazer entregas regulares e antecipadas de incrementos de *software,* de modo que os interessados obtenham maior vantagem competitiva possível [Agile Alliance 2017].

Com isso, neste artigo, o método do VPL foi escolhido como função objetivo para ser aplicado e avaliado sob duas diferentes abordagens em um projeto hipotético de desenvolvimento iterativo de *software.* A primeira abordagem propõe o emprego do método sob a ótica do *Critical Path Method* (CPM) (sem restrição de recursos, mas respeitando a relação de precedência entre atividades ou módulos) e a segunda abordagem propõe o emprego do método sob a ótica de [Denne e Cleland-Huang 2003] (com restrição de recursos e respeitando a relação de precedência entre atividades ou módulos).

2. Referencial Teórico

a. VPL como função objetivo

Avaliações econômicas e financeiras de um projeto podem ser realizadas por diferentes métodos, que podem ser utilizados de forma combinada ou isolada. Tais métodos, também podem ser chamados de função objetivo, pois refletem aspectos de interesse de seus patrocinadores. Embora existam vários métodos passíveis de aplicação em projetos de *software,* neste artigo, apenas o VPL foi utilizado e avaliado.

Na prática, o VPL é um instrumento usado para determinar o valor presente equivalente a pagamentos futuros. Deste modo, desconta uma taxa de juros adequada, em função do que normalmente é chamado de Taxa Mínima de Atratividade (TMA). Para encontrar o VPL é necessário conhecer o investimento inicial, o fluxo de caixa líquido em uma data específica e o custo de capital definido pelos interessados. A fórmula (1) permite calcular o VPL, onde *n* é cada um dos períodos sob avaliação (e varia do período 1 até o período *N*), *FCt* é um fluxo de caixa em um período de tempo e *i* é a taxa a ser descontada (custo do capital).

$$\textbf{VPL} = \sum_{n=1}^{N} \frac{FC_t}{(1+i)^n} \quad (1)$$

Em alto nível, pode ser dito que quando o valor do VPL é maior que 0 o projeto é lucrativo, quando o valor do VPL é exatamente 0 o projeto consegue pagar seus custos e, por último, quando o valor do VPL é menor que 0 o projeto traz prejuízo [Shim e Siegel 2001].

b. Abordagem *Critical Path Method* (CPM)

O CPM é um método utilizado em projetos de diversos segmentos de negócio para sequenciamento de atividades ou módulos de projetos [Kerzner 2001]. Sua utilização define uma rede de relações que destaca o conjunto de precedência entre atividades ou módulos de um projeto. Uma das notações possíveis para o CPM considera que os nós são as atividades ou módulos e as arestas são as relações de precedência entre os nós. Conhecendo a duração dos nós é possível identificar o caminho crítico e as eventuais folgas no projeto. A figura 01

apresenta um exemplo de diagrama utilizado para avaliação do método de CPM. Nesta figura, assumindo que as caixas A, B, C, D, E e F (nós que representam módulos ou atividades) têm duração de duas unidades de tempo e todas têm relacionamentos do tipo término-término, diz-se que apenas o subconjunto de caixas A, B, C e F representa o caminho crítico da rede, pois não pode sofrer qualquer atraso sem comprometer o final do projeto. Já as caixas D e E são caixas fora do caminho crítico, pois podem sofrer algum atraso sem necessariamente comprometer o fim do projeto.

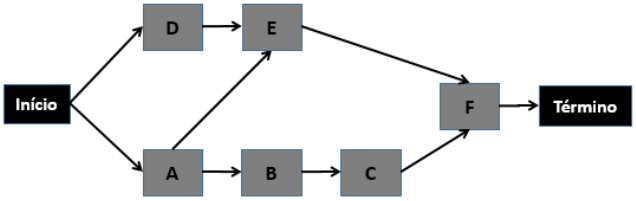


Figura 01. Diagrama de Precedência com Simultaneidades

O método do CPM considera que as caixas que não mantêm relacionamento entre si podem ser programadas o quanto antes. Por exemplo, as caixas A e D não mantêm relação entre si e, portanto, estão programadas em paralelo logo no início do projeto. Isso pode ser um problema se não houver recurso suficiente para a execução de todas as caixas. Se, eventualmente, a capacidade de execução da equipe for de uma atividade por vez, a programação da figura 01 se torna inviável. Assim, no método de CPM, não há preocupação com o limite de recursos, pelo menos inicialmente.

c. Abordagem Denne e Cleland-Huang

[Denne e Cleland-Huang 2003] encorajam a decomposição de um *software* em unidades capazes de gerar valor para o negócio. Essas unidades devem ser implementadas em pequenos ciclos de desenvolvimento, favorecendo o fluxo de caixa de interesse para o projeto. Tais unidades são chamadas de *Minimum Marketable Feature Modules* (MMF) e criam valor em pelo menos uma das seguintes áreas de uma organização: vantagem competitiva, geração de receita, economia de custos, projeção de marca e fidelização de clientes.

Embora MMFs sejam consideradas unidades independentes, Denne e Cleland-Huang também apresentam o conceito de *Architectural Elements* (AE), um tipo de unidade necessário para apoiar as MMFs. No entanto, um AE não oferece valor direto aos interessados no projeto. Deste modo, a decomposição deve ser feita em termos de MMFs e AEs, a fim de definir o arranjo com o melhor retorno [Alencar; Franco e Schmitz 2013].

A abordagem de Denne e Cleland-Huang considera ainda que o desenvolvimento das MMFs e AEs deve respeitar tanto as respectivas relações de precedência entre as unidades, bem como o limite de capacidade de uma unidade por vez. Em outras palavras, duas unidades (MMFs ou AEs) não podem ser desenvolvidas simultaneamente, mesmo que não mantenham relação entre si, o que é diferente da abordagem anteriormente assumida para a aplicação de CPM. A figura 02 leva em consideração o mesmo número de caixas (nós) da figura 01 e também as mesmas relações entre as caixas (nós), no entanto, assume-se que só há capacidade para execução de uma caixa por vez. Desta maneira, a figura 02 apresenta deslocamentos a fim de eliminar simultaneidades de nós. Esse tipo de deslocamento costuma aumentar o tempo de execução do conjunto de nós, mas dilui os gastos em um maior período.

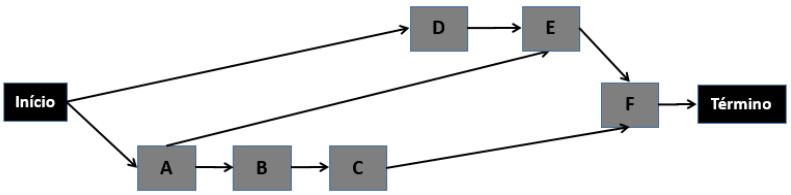


Figura 02. Diagrama de Precedência sem Simultaneidades

3. Metodologia

a. Tipo de Pesquisa

Para [Vergara 2007], uma pesquisa pode ser classificada quanto aos fins, revelando seus objetivos, e quanto aos meios, revelando os métodos a serem utilizados. Deste modo, quanto aos fins o trabalho é: Descritivo, pois expõe características de um projeto de *software*; Explicativo, pois visa esclarecer diferentes resultados, por meio da função objetivo VPL, utilizada sob as abordagens de CPM e de [Denne e Cleland-Huang 2003]; Aplicado, pois tem por motivação resolver problemas concretos. Já quanto aos meios, o trabalho pode ser classificado como: Bibliográfico, pois está baseado em conceitos já publicados; Pesquisa de laboratório, pois inclui simulações de distribuição de MMFs e AEs, através de algoritmos desenvolvidos em R [R 2007] e também é Estudo de caso, pois se refere a avaliações variadas do mesmo problema sobre um único projeto hipotético de desenvolvimento de *software* (dando caráter de aprofundamento).

b. Cenário Hipotético do Estudo

Para o desenvolvimento das análises e comparações entre as abordagens de CPM (sem restrição de recursos) e de Denne e Cleland-Huang (com restrição de recursos), o diagrama de precedência da figura 03 será usado como o cenário de avaliação do projeto hipotético. Neste diagrama, os nós 1 e 11 são do tipo *Dummy* (nó que marca o início ou o fim de uma rede de precedência), os nós 3, 4, 5, 7, 8 e 10 são do tipo MMF (nó que representa menor unidade comercializável de funções, com capacidade de retorno financeiro) e os nós 2, 6 e 9 são do tipo AE (estrutura arquitetônica que não promove retorno financeiro diretamente, mas apoia MMF em seu potencial de retorno).

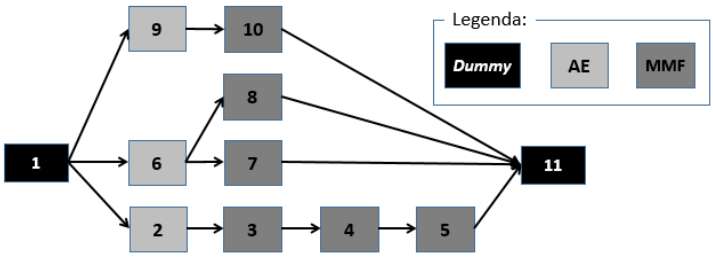


Figura 03. Diagrama de Precedência do Projeto Hipotético

Para esse mesmo projeto, a tabela 01 (indicada no apêndice A) apresenta os respectivos fluxos de caixa para cada uma das unidades (MMFs e AEs). Nesta tabela, são apresentados o ID da unidade, o tipo de unidade, a duração, as unidades predecessoras e os valores (positivos ou negativos) para cada um dos períodos. Na tabela 01, com uma janela de avaliação de 16 períodos, todas as unidades estão programadas para a data mais cedo, ou seja, as restrições de precedência da figura 03 não foram aplicadas ainda. Além disso, os valores não acumulados e acumulados se referem a valores nominais (fluxo sem desconto de taxa).

As duas próximas seções apresentam as distribuições para tais unidades (MMFs e AEs), considerando as restrições de precedência da figura 03, nas perspectivas de CPM e Denne e Cleland-Huang.

4. VPL via abordagem do CPM

O VPL, como descrito anteriormente, é uma função objetivo utilizada para indicar o valor presente que equivale a pagamentos futuros em um projeto. Desta maneira, uma taxa de desconto deve ser considerada para definir a equivalência. Já o método do CPM, identifica as restrições de precedência entre as atividades (chamadas aqui de unidades MMFs ou AEs), de modo a estabelecer o sequenciamento para um projeto. Como é possível ter várias sequências válidas, diante das restrições da figura 03, entre as unidades (MMFs e AEs), foi elaborado um algoritmo em R capaz de identificar todas as sequências possíveis e seus respectivos fluxos de caixa nominais e descontados. Esse algoritmo foi capaz de identificar 2621 sequências diferentes para as MMFs e AEs, respeitando as restrições de precedência do projeto hipotético. Desse conjunto, o melhor VPL encontrado está indicado na tabela 02 (também no apêndice A), que apresenta o ID da unidade, tipo (MMF ou AE), duração, unidades predecessoras e também o fluxo de caixa dentro da janela de 16 períodos. A tabela 02 apresenta ainda a taxa de desconto utilizada (2,41 % a.a.), os fluxos nominais (não acumulado e acumulado - AC) e os fluxos descontados (não acumulado e acumulado – AC).

A aplicação das restrições indicadas na figura 03 faz com que as MMFs de IDs 3, 4, 5, 7, 8 e 10 tenham algum deslocamento para a direita, quando comparadas aos dados da tabela 01 (que desconsideram as restrições). Como consequência, isso faz com que parte dos fluxos de caixa dessas MMFs ultrapassem a janela de oportunidade de 16 períodos pré-definida. Deste modo, o VPL acumulado ao final dos 16 períodos do projeto tem um valor de $6.926,20. Em valor nominal (sem desconto), o fluxo de caixa acumulado indica o valor de $9.738,00.

Já o pior VPL encontrado pelo algoritmo, do universo de 2621 sequências, foi o de $4.737,60. Ou seja, a diferença entre o melhor VPL e o pior VPL foi de $1.558,26 (6.926,20 – 4.737,60), uma queda de cerca 24% do primeiro para o segundo valor. Do total de sequências identificadas pelo algoritmo, 419 (cerca de 16% de todas as sequências) apresentaram VPL variando de $5.901,00 a $6.926,20 (cerca de 85% a 100% do valor do melhor VPL). Por outro lado, cerca de 2200 sequências (cerda de 84% de todas as sequências) têm VPL variando de $4.737,60 a $5.900,00 (68% a 85% do valor do melhor VPL). Isso demonstra que poucas sequências identificadas pelo algoritmo concentraram os melhores valores de VPL.

5. VPL via abordagem Denne e Cleland-Huang

Baseado no estudo de [Denne e Cleland-Huang 2003], a função objetivo VPL também foi avaliada para o projeto da figura 03. No entanto, além das restrições da figura 03, também se considerou que só é possível trabalhar em uma unidade (MMF ou AE) de cada vez. Em outras palavras, em função de restrição de recursos, cada unidade deve ser realizada sem paralelismo com outra. A utilização de um algoritmo elaborado em R permitiu identificar mais de 2500 sequências possíveis para o projeto da figura 03, levando em consideração apenas uma unidade por vez. De todas as sequências, a que apresentou o melhor VPL está demonstrada através da tabela 03 (indicada no apêndice A). Nesse caso, o valor de melhor VPL foi de $2.992,50. Também é importante dizer que, de todas as sequências possíveis, o pior VPL foi o de valor $658,50 (um valor quatro vezes menor do que o maior VPL).

6. Análise e Discussão

As simulações de CPM identificaram VPL máximo no valor de $6.926,20, enquanto as de Denne e Cleland-Huang de $2.992,50 (menos da metade do 1º valor). Quanto ao menor VPL, as simulações de CPM identificaram o valor de $4.737,60, já as de Denne e Cleland-Huang identificaram o valor de $658,50 (quase 7 vezes menos). Isso demonstra que a abordagem de CPM além de identificar um VPL máximo muito melhor também mostra que a distância entre o seu melhor e pior VPL é menor. O pior VPL de CPM representa cerca de 69% do seu melhor VPL. Já o pior VPL de Denne e Cleland-Huang representa apenas 22% do seu melhor VPL.

Além disso, do total de sequências identificadas (2621) pelo algoritmo sem restrição de recurso, 84% não superaram os 85% do valor do melhor VPL. Ou seja, apenas 16% das sequências foi capaz de gerar VPL a partir de 85% do melhor VPL. Quanto às sequências identificadas pelo algoritmo sob a abordagem de Denne e Cleland-Huang (cerca de 2500 encontradas), 99,5% não foram capazes de superar o valor de 85% do melhor VPL. Ou seja, apenas 0,5% de todas as sequências de Denne e Cleland-Huang foram capazes de atingir 85% ou mais do melhor VPL. Isso demonstra que existem mais opções de maximização de VPL com a abordagem utilizada para CPM (sem restrição de recursos) do que para a abordagem de Denne e Cleland-Huang. Embora fosse esperado que a abordagem de CPM promovesse as melhores opções de VPL, sua aplicação envolve maior mobilização de gastos no projeto, nos pontos em que há paralelismo. A antecipação das unidades com CPM faz com que os benefícios comecem a ser auferidos mais cedo (aumentando o VPL). Por outro lado, existe um preço a pagar, pois o fluxo de caixa de despesas ao invés de ser diluído em um maior período, passa a ter maiores concentrações no início, podendo trazer problemas de financiamento para o projeto, como pode ser visto na figura 04, entre os períodos de 1 a 6 (pelo eixo horizontal).

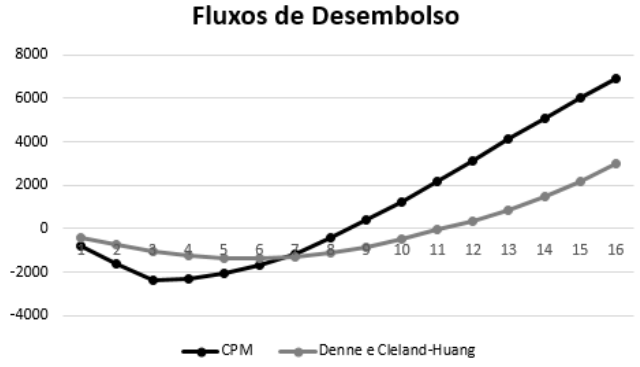


Figura 04. Fluxos de Desembolso

7. Conclusões

a. Considerações Finais

A principal diferença entre as abordagens utilizadas para a avaliação da função objetivo deste trabalho foi a condição de uso de recursos. Ou seja, na abordagem de CPM, considerou-se que não havia restrição de recursos para a realização de unidades (MMFs e AEs) simultaneamente, desde que respeitadas as restrições de precedência. Já na abordagem de Denne e Cleland-Huang, considerou-se que os recursos só teriam capacidade de realização de uma unidade por vez (MMFs e AEs), além de também respeitar as restrições de precedência. Deste modo, a abordagem de CPM acabou superando a abordagem de Denne e Cleland-Huang, embora com maior exigência de investimento em momentos iniciais do projeto.

b. Sugestões de Estudos Futuros

A avaliação individual e combinada de funções objetivo em projetos de desenvolvimento de *software* não se esgota com as análises feitas neste artigo. Várias vertentes e aspectos podem ser investigados, de modo a aprofundar e favorecer o processo de tomada de decisão em projetos de *software*. Entre as possibilidades, estão:

- Se os principais interessados de um projeto de *software* ou de uma carteira de projetos de *software* têm mais de uma função objetivo a considerar no processo de tomada de decisão, como essas funções podem ser ponderadas, de modo a favorecer a identificação das melhores opções de execução? Ou seja, como definir a melhor estratégia de decisão, de acordo com as funções objetivo selecionadas, e que avaliação relativa as funções devem manter entre si?
- A identificação das melhores opções de execução para um projeto de *software* ou uma carteira de projetos de *software*, muitas vezes, tem natureza heurística, ou seja, a investigação é baseada na aproximação progressiva de uma solução. Levando em consideração essa premissa, é possível se beneficiar de lógicas não clássicas como Lógica *Fuzzy* ou Lógica Paraconsistente na identificação de soluções aproximadas, já que essas lógicas admitem opções de resultado além de somente "Falso" ou "Verdadeiro", como por exemplo, "Quase Verdadeiro" ou "Quase Falso"? Ou seja, de que maneira lógicas não clássicas podem favorecer a identificação de soluções aproximadas no contexto de tomada decisão para projetos (talvez com métodos como Monte Carlo) diante de diferentes funções objetivo?

7. Referências

AGILE ALLIANCE. (2017). **Manifesto for Agile Software Development.** Disponível em <http://www.agilemanifesto.org>. Acesso em Maio, 2017.

ALENCAR, J. A.; FRANCO, C. A. S.; SCHMITZ, E. A.; CORREA, A. L. (2013). *A statistical approach for the Maxiization of the Financial Benefits Yielded By a Large Set of MMFs and AEs. Computing and Informatics,* vol.32.

DENNE, M.; CLELAND-HUANG, J. (2003). **Software by Numbers: Low-Risk, High-Return Development.** Sun Microsystems.

IEEE (1993). *Standards Collection: Software Engineering, IEEE Standard 610.12-1990,* IEEE.

JACOBSON, I.; RUMBAUGH, J.; BOOCH, G. (1999). **Unified Softwar Developement Process.** Addison-Wesley, Reading – MA.

KERZNER, H. (2001). **Project Management: A Systems Approach to Planning, Scheduling and Controlling.** 7th Edition. Wiley.

NAUR, P.; RANDALL, B. (1969). *Software Engineering: A Report on a Conference Sponsored by the NATO Science Committee,* NATO.

PRESSMAN, R. S. (2006). **Engenharia de Software.** 6ª Edição. McGraw Hill – São Paulo.

R (2017). Disponível em http://cran.r-project.org. Acesso em 06/09/2017.

SHIM, J. K.; SIEGEL, J. G.(2001). *Handbook of Financial Analysis, Forecasting, and Modeling.* 2th.Ed.

VERGARA, S. C. (2007). **Projetos e Relatórios de Pesquisa em Adm.** 9ª Ed. Atllas. S.Paulo.

8. Apêndice A – Tabelas de Fluxos de Caixa

Tabela 01. Fluxos de Caixa para MMFs e AEs (todos os nós no início)

ID	Tipo	Duração	Precessor	Custo ou Receita por Período ($)															
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	MMF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	AE	1	1	-200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	MMF	1	2	-200	100	100	90	80	70	60	50	40	30	20	10	0	0	0	0
4	MMF	2	3	-200	-200	100	130	160	190	220	250	250	250	250	250	250	250	250	250
5	MMF	3	4	-200	-200	-100	140	180	220	260	300	340	380	400	400	400	400	400	400
6	AE	1	1	-400	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	MMF	2	6	-250	-250	50	80	100	120	140	160	18	200	200	200	200	200	200	200
8	MMF	2	6	-350	-350	50	100	150	200	250	300	350	350	350	350	350	350	350	350
9	AE	1	1	-200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	MMF	2	9	-100	-100	100	100	150	150	150	150	150	150	150	150	150	150	150	15
11	MMF	0	5, 7, 8, 10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Valores não acumulados				-2100	-1000	300	640	820	950	1080	1210	1148	1360	1370	1360	1350	1350	1350	1215
Valores acumulados				-2100	-3100	-2800	-2160	-1340	-390	690	1900	3048	4408	5778	7138	8488	9838	11188	12403

Tabela 02. Melhor VPL CPM

ID	Tipo	Duração	Precessor	Custo ou Receita por Período ($)															
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	MMF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	AE	1	1	-200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	MMF	1	2		-200	100	100	90	80	70	60	50	40	30	20	10	0	0	0
4	MMF	2	3			-200	-200	100	130	160	190	220	250	250	250	250	250	250	250
5	MMF	3	4					-200	-200	-100	140	180	220	260	300	340	380	400	400
6	AE	1	1	-400	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	MMF	2	6		-250	-250	50	80	100	120	140	160	18	200	200	200	200	200	200
8	MMF	2	6		-350	-350	50	100	150	200	250	300	350	350	350	350	350	350	350
9	AE	1	1	-200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	MMF	2	9		-100	-100	100	100	150	150	150	150	150	150	150	150	150	150	15
11	MMF	0	5, 7, 8, 10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fluxo nominal				-800	-900	-800	100	270	410	600	930	1060	1028	1240	1270	1300	1330	1350	1350
Fluxo nominal (AC)				-800	-1700	-2500	-2400	-2130	-1720	-1120	-190	870	1898	3138	4408	5708	7038	8388	9738
Fluxo descontado				-781,174	-858,139	-744,84	90,914	239,691	355,41	507,872	768,677	855,509	810,157	954,236	954,323	953,877	952,925	944,492	922,265
Fluxo desconotado (AC)				-781,174	-1639,31	-2384,15	-2293,24	-2053,55	-1698,14	-1190,27	-421,588	433,921	1244,08	2198,31	3152,64	4106,51	5059,44	6003,93	6926,2
Taxa % a.a.	2,41																		

Tabela 03. Melhor VPL Denne e Cleland-Huang

ID	Tipo	Duração	Precessor	Custo ou Receita por Período ($)															
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	MMF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	AE	1	1						-200	0	0	0	0	0	0	0	0	0	0
3	MMF	1	2							-200	100	100	90	80	70	60	50	40	30
4	MMF	2	3								-200	-200	100	130	160	190	220	250	250
5	MMF	3	4										-200	-200	-100	140	180	220	260
6	AE	1	1	-400	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	MMF	2	6				-250	-250	50	80	100	120	140	160	18	200	200	200	200
8	MMF	2	6		-350	-350	50	100	150	200	250	300	350	350	350	350	350	350	350
9	AE	1	1													-200	0	0	0
10	MMF	2	9														-100	-100	100
11	MMF	0	5, 7, 8, 10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fluxo nominal				-400	-350	-350	-200	-150	0	80	250	320	480	520	498	740	900	960	1190
Fluxo nominal (AC)				-400	-750	-1100	-1300	-1450	-1450	-1370	-1120	-800	-320	200	698	1438	2338	3298	4488
Fluxo descontado				-390,59	-333,72	-325,87	-181,83	-133,16	0	67,716	206,63	258,27	378,28	400,16	374,21	542,98	644,84	671,64	812,96
Fluxo desconotado (AC)				-390,59	-724,31	-1050,2	-1232	-1365,2	-1365,2	-1297,4	-1090,8	-832,55	-454,26	-54,101	320,11	863,09	1507,9	2179,6	2992,5
Taxa % a.a.	2,41																		

DICTA: Biblioteca para reconhecimento de elocuções baseada em uma rede neural sem peso

Ericson J. S. Soares[1], Diego F. P. de Souza[2], Priscila M. V. Lima[3]

[1]Departamento de Ciência da Computação
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

[2]Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa em Engenharia (COPPE)
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

[3]Instituto Tércio Pacitti (NCE)
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

ericson@dcc.ufrj.br, diegosouza@cos.ufrj.br, priscila.lima@nce.ufrj.br

Abstract. *The task of efficiently recognizing utterances, that is, sounds emitted by the human voice is still a research topic in the present times. DICTA is an open source project for the recognition of vocal expressions based on a weightless neural network model, more specifically the WiSARD (Wilkes, Stonhan and Aleksander Recognition Device). Its ultimate goal is to become a general purpose library to the recognition of any set of vocal utterances from a single user, being able to be integrated into digital inclusion systems for the visually impaired and quadriplegic, for example.*

Resumo. *A tarefa de reconhecimento eficiente de elocuções, isto é, sons emitidos pela voz humana ainda é um tópico de pesquisa nos tempos atuais. DICTA é um projeto de código aberto para o reconhecimento de expressões vocais baseado num modelo de rede neural sem peso, mais especificamente na WiSARD (Wilkes, Stonhan e Aleksander Recognition Device). Seu objetivo final é se tornar uma biblioteca de uso geral para o reconhecimento de qualquer conjunto de expressões vocais de um único usuário, podendo ser integrada em sistemas de inclusão digital para deficientes visuais e tetraplégicos, por exemplo.*

1. Introdução

Elocuções constituem sons emitidos pela voz humana, tais como palavras, letras ou sons mais rudimentares. Atualmente existem várias aplicações que se beneficiariam da disponibilidade de uma biblioteca para o reconhecimento de elocuções. Uma aplicação bastante relevante desse tipo de biblioteca seria em sistemas de inclusão digital para deficientes visuais e tetraplégicos, embora exista aplicabilidade em inúmeros sistemas de informação. Resultados preliminares obtidos por colaboradores [1] [2] indicaram que se fazem necessárias três etapas distintas no sistema para o processamento e para o reconhecimento de áudio: (i) pré-processamento do sinal de áudio [3], (ii) normalização do sinal processado para uma entrada de tamanho padrão e (iii) o treinamento / classificação

do sinal em si. O pré-processamento do áudio inclui manipulações matemáticas das ondas do sinal digital, tais como Transformada de Fourier e mudança na escala de frequência das ondas para uma forma que mais se aproxime do modo como são reconhecidas pelo ouvido humano. Desse modo, destacam-se as características de maior importância em cada exemplar de som a ser reconhecido. A normalização aplicada neste trabalho utiliza a técnica de *KernelCanvas*[1] [2], que converte um sinal de um tamanho qualquer para um tamanho padrão. Para a etapa de treinamento e reconhecimento foi explorado um modelo de aprendizado de máquina para classificação baseado em N-Tuplas, mais precisamente o classificador *WiSARD* [4] [5]. Essa escolha foi feita dadas as características do classificador, que permite aprendizado *online* e possui baixa complexidade computacional. Isso permite retreinar o reconhecedor de maneira rápida em caso de mudanças de voz ou de capacidade de articulação dos mesmos.

O objetivo principal desta etapa da pesquisa foi o desenvolvimento de uma biblioteca que conseguisse processar arquivos de áudio no formato *WAV* com elocuções distintas de um único usuário e classificá-las corretamente. A partir dessas classificações, deseja-se validar o uso da rede neural sem peso *WiSARD* para essa tarefa. Isto permitirá que, futuramente, a biblioteca seja estendida para processar e classificar em tempo real o áudio capturado diretamente de um microfone conectado ao computador.

2. Biblioteca DICTA: Principais conceitos e etapas

Conforme já mencionado na parte introdutória, a metodologia aplicada no desenvolvimento da biblioteca dividia a mesma em três etapas distintas. A primeira delas consistia do pré-processamento do sinal de áudio, seguida da etapa de normalização do sinal processado para uma entrada de tamanho padrão e, por último, da realização do treinamento ou classificação do sinal.

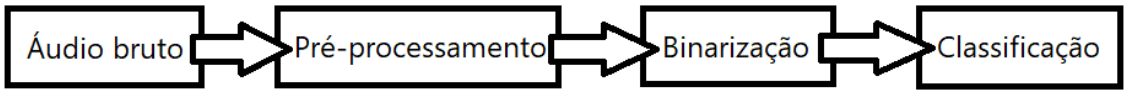


Figura 1. Etapas de funcionamento da biblioteca

Para fins de reconhecimento de fala, porém, existem alguns problemas que devem ser considerados e que podem prejudicar os resultados. O sinal gravado normalmente apresenta ruído, ou seja, sons emitidos por outras fontes que não a voz em questão. A duração e o volume de sinais que representam uma mesma classe de elocuções podem variar. Além disso, ainda há particularidades inerentes ao locutor, como diferença da proporção de agudos e graves, diferentes frequências base da voz, ou particularidades resultantes do processo natural de fala, desde diferentes durações dos fonemas até problemas degenerativos no aparelho fonador. Com o intuito de amenizar essas questões inerentes da captura de voz humana a partir de um microfone, se fazem necessárias as etapas de pré-processamento descritas nesta seção.

2.1. Pré-processamento

O principal papel do pré-processamento é transformar um áudio bruto (captado diretamente de um microfone e gravado em arquivo *WAV*) e destacar informações relevantes

sobre ele tendo em vista o objetivo de reconhecer a fala humana. O primeiro procedimento a se fazer no tratamento do áudio é separá-lo em trechos de curta duração, em torno de 20 a 40 ms [3]. Esse processo é chamado de *framing* e é de suma importância para o restante do processamento. Como o áudio gravado passa por diversas mudanças ao longo do tempo (diferentes fonemas são falados), é mais interessante tratar cada pedacinho ”constante” individualmente e analisar as transições entre eles. O *framing* divide o áudio todo em *frames* de mesmo tamanho. Os *frames* são dispostos de forma que haja sobreposição entre eles, facilitando a detecção e inclusão das transições. Após a conclusão do *framing*, executa-se a etapa de *windowing*. O *windowing* consiste em passar as amplitudes de cada frame por uma máscara de intensidade. O objetivo é amenizar as amplitudes das extremidades, de modo a melhorar o resultado do passo seguinte, a Transformada de Fourier. Existem diversas funções capazes de realizar esse procedimento. A escolhida para o projeto foi a função de Hann [6]. Essa função tem o valor zero em suas extremidades e cresce até 1 no meio do intervalo de forma suave. Isso significa que o meio de cada *frame* é mantido praticamente igual ao original, enquanto as partes mais próximas das extremidades tem diminuída a intensidade de seus valores, que se aproximam de zero. Por não se tratar de um sinal contínuo, devemos aplicar a versão discreta da Transformada de Fourier (DFT) [3]. O algoritmo utilizado para tal foi o *Fast Fourier Transform* (FFT), seu resultado é igual ao do DFT, mas sua complexidade é consideravelmente mais baixa, passando de $O(n^2)$ do algoritmo original para $O(nlog(n))$. Em um sinal sonoro, os coeficientes da Transformada de Fourier podem ser vistos como a contribuição de cada frequência para a composição do sinal.

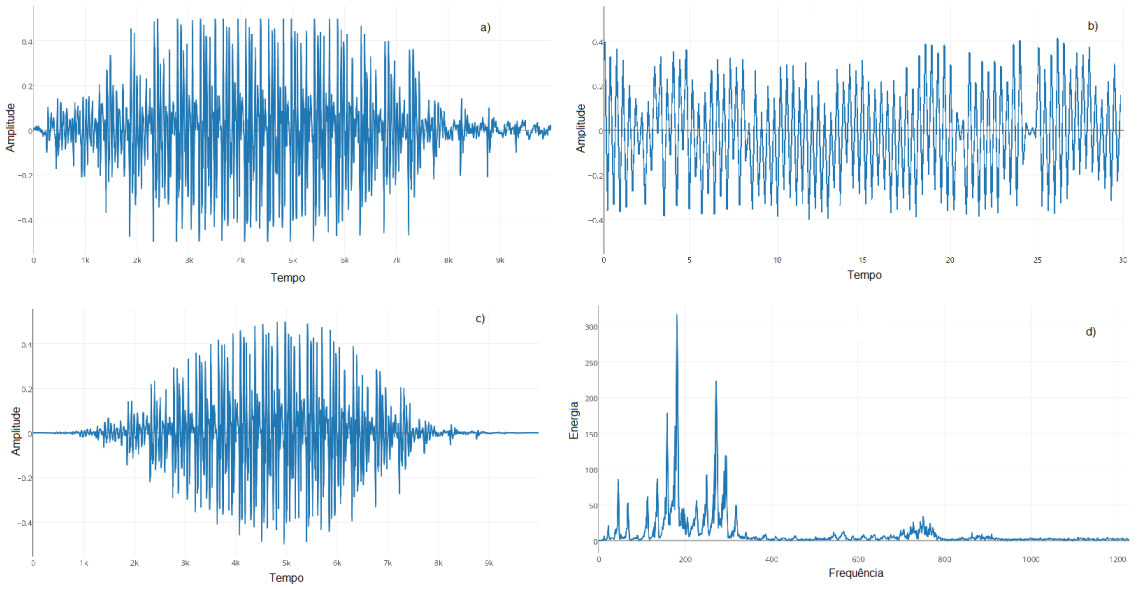


Figura 2. Nessa figura temos: a) Áudio original; b) Um *frame* isolado desse áudio; c) O áudio original após o processo de *windowing*; d) Após a Transformada de Fourier

Uma vez tendo as intensidades separadas por frequências, fica mais fácil destacar as frequências importantes para o reconhecimento de voz. A técnica escolhida para tratar essa parte do pré-processamento foi a técnica dos *filter banks*. Para que isso seja possível, entretanto, ainda é preciso realizar alguns passos intermediários. O primeiro deles é con-

verter as frequências de Hertz para Mels [3]. A conversão se dá de forma extremamente simples, seguindo a fórmula (1), onde f é a frequência em Hertz e o resultado está em Mels.

$$M(f) = 1125ln(1 + f/700) \tag{1}$$

Para reverter de Mels para Hertz, basta aplicar a função inversa, descrita em (2), sendo m o valor em Mels.

$$M^{-1}(m) = 700(\exp(m/1125) - 1) \tag{2}$$

A grande vantagem de trabalhar em Mels é que ela é uma escala logarítmica com relação a escala Hertz. Isso representa de forma mais confiável o padrão de reconhecimento auditivo humano. Humanos têm a tendência de reconhecer fenômenos (auditivos ou visuais) em escala logarítmica. E essa mudança permite destacar frequências importantes na voz humana. Com a escala similar à presente na percepção humana, agora, 26 valores são distribuídos linearmente entre limites empiricamente determinados. A frequência mais baixa considerada é 300Hz e a mais alta 8000Hz (intervalo sobre o qual a voz humana normalmente atua, tendo em vista que a frequência de gravação deve ser o dobro da original). Os valores limitantes são convertidos para Mels e assim, a distribuição é feita. Por ser uma distribuição linear em Mels, quando convertidas de volta para Hertz, a distribuição será logarítmica, exatamente como o esperado. A quantidade de 26 *filter banks* foi empiricamente estabelecida para melhores resultados de reconhecimento [3].

Finalmente, a importância das frequências será alterada. Como dito anteriormente, isso é feito com a técnica de *filter banks* [3]. Essa técnica consiste em usar os 26 valores escolhidos no passo acima como únicas frequências para descrever o *frame* em questão. Para cada *filter bank* (um dos 26 valores), atribui-se um intervalo de atuação, intervalo esse que inclui a frequência do *filter bank* em questão. Cada *filter bank* terá um único valor, resultante da média ponderada das frequências presentes dentro de seu intervalo. Os pesos de cada frequência são determinados de acordo com a sua proximidade à frequência do *filter bank* em questão. Isso é feito de forma que quanto mais próxima estiver a frequência de um *filter bank*, mais próximo o peso será de 1. Os pesos diminuem linearmente de acordo com o afastamento da frequência do *filter bank*, onde as extremidades do intervalo são iguais ao valor zero. Após os 26 *filter banks* terem sido calculados para cada *frame* do áudio original, o último passo do pré-processamento é calcular o cepstrum [3] desses *filter banks* e normalizá-los. Para isso, tira-se o logaritmo natural de cada *filter bank* e, em seguida, esses novos valores são usados para calcular a Transformada Discreta de Cossenos (DCT). Para concluir a normalização, cada resultado é substituído pela soma de todos os pontos anteriores. A seguir, a média e o desvio padrão desses pontos são calculados e, por fim, subtraímos a média de cada um dos pontos e dividimos pelo desvio padrão. Finalmente aplicamos a tangente hiperbólica nos pontos e o resultado estará normalizado entre -1 e 1.

2.2. Binarização

Agora, o arquivo original já foi tratado e suas características importantes estão armazenadas em vetores de 26 posições, um para cada *frame*, encerrando assim a etapa de pré-processamento. Esse sinal pode ser interpretado como um caminho, ou traçado, em $\mathbb{R}^{26}$, onde cada ponto corresponde a um *frame*, gerado durante o pré-processamento. Como o classificador do projeto aceita apenas uma entrada binária, é necessário usar um

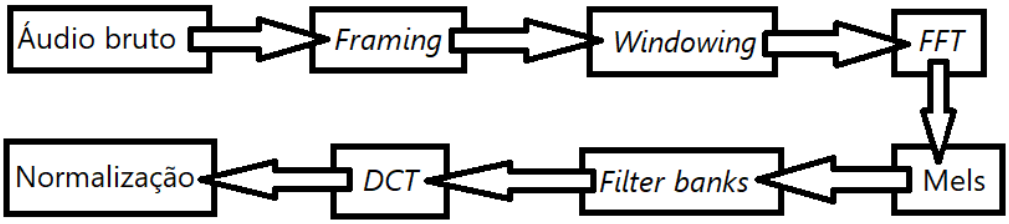


Figura 3. Etapas do pré-processamento

algoritmo capaz de traduzir esse padrão multi-dimensional real em algo binário. Para resolver essa questão, usamos o *KernelCanvas* [1][2].

O *KernelCanvas* é uma técnica de discretização binária multi-dimensional. Primeiramente, um número arbitrário de pontos n-dimensionais, a partir de agora chamados de *kernels*, é escolhido. Cada *kernel* fica responsável por uma região do espaço na qual cada ponto dentro dela tem como *kernel* mais próximo o *kernel* responsável pela região. [1][2]. Cada um dos *kernels* tem um valor, ativado ou não ativado. Um *kernel* é considerado ativado se e somente se pelo menos um dos pontos resultantes do pré-processamento está dentro de sua região. Caso contrário, ele é considerado não ativado. O resultado do *KernelCanvas*, então, disponibiliza uma lista de valores binários, exatamente o que é necessário para usar como entrada no classificador, como cada *kernel* representa N *bits* no padrão binário da saída, caso este *kernel* esteja marcado como ativo, todos os N pontos terão valor 1, do contrário todos terão valor 0, ou vice-versa.

Resta saber, então, como escolher as coordenadas dos *kernels*. Os *kernels* podem ser distribuídos no espaço de maneira aleatória ou distribuídos no espaço de forma a ficarem igualmente espaçados uns dos outros, ou então uma mistura dessas duas distribuições, onde temos *kernels* com espaçamentos razoavelmente similares uns dos outros. A versão implementada, por recomendação, foi a distribuição puramente aleatória [1][2]. Outro fator importante é o intervalo de sorteio dos *kernels*. Como a saída do pré-processamento foi normalizada para ficar contida entre -1 e 1, esses são os limites em cada dimensão para o sorteio dos *kernels*.

2.3. Classificação

Com isso a entrada já está tratada, binarizada e pronta para ser utilizada na rede neural *WiSARD*, o classificador. A *WiSARD* é um algoritmo leve e simples de classificação que possui apenas dois modos de operação [4][5]. O primeiro é o de treinamento, na qual ela aprende de forma supervisionada a classificar suas entradas. O segundo é o de teste, ou classificação, no qual a rede classifica a entrada baseando-se no treinamento que teve. O algoritmo pode ser pensado como uma junção de partes separadas.

A primeira parte é chamada de retina. A retina nada mais é do que um vetor unidimensional binário. Ou seja, cada posição do vetor pode assumir 0 ou 1 e esse vetor é uma lista contínua de tamanho arbitrário. Para fins semânticos, é possível visualizar este vetor unidimensional como vetores n-dimensionais que tenham alguma relevância de significado com o problema. Por exemplo, se a entrada for uma imagem, a retina é um vetor bidimensional e cada posição representa um pixel. Para o algoritmo isso realmente não importa e quando se trata de dados em dimensões mais altas, essa visualização se perde.

No caso desse projeto, a entrada é uma sequência de *bits*, uns ou zeros, que representam os *kernels* marcados como ativados ou não ativados. Sobre a saída do *KernelCanvas* (entrada da *WiSARD*): Nesse relatório a retina será visualizada como um vetor unidimensional. A segunda parte consiste em criar os chamados discriminadores. Discriminadores são os responsáveis por detectar certos tipos de padrões relacionados a determinada classe. Assim, normalmente se cria um discriminador por classe e cada discriminador fica responsável por detectar padrões de sua classe. Dessa forma, quando em modo de treinamento, a rede ensinará ao discriminador que características a retina deve ter para que seja classificada como de sua classe. Já em modo de teste, o discriminador avaliará se a retina tem as características que ele aprendeu durante o treinamento e dirá o quão similar é a entrada em relação a sua classe.

Para que os discriminadores sejam capazes de aprender precisamos dar a eles diversas *RAMs* [4][5]. As *RAMs* são pequenas memórias que guardam características vistas na retina e as associam à classe correspondente por meio do discriminador. Cada *RAM* fica responsável por uma parte, geralmente não contínua, da retina e guarda informações a respeito de sua classe apenas levando em consideração sua parte. Se juntarmos todas as *RAMs* de um dado classificador, temos a retina inteira. Por exemplo, suponha o exemplo da Figura 4 onde temos uma *RAM* responsável pela segunda, quarta, e sexta posições da retina. Nesse caso a *RAM* leria 1 da posição dois, 1 da posição quatro, e 0 da posição seis. O que a *RAM* faz, quando em estado de treinamento, é concatenar esses valores, formando 110, e marcar o valor 1 na posição de memória do índice 6, que corresponde ao valor decimal do endereço binário 110.

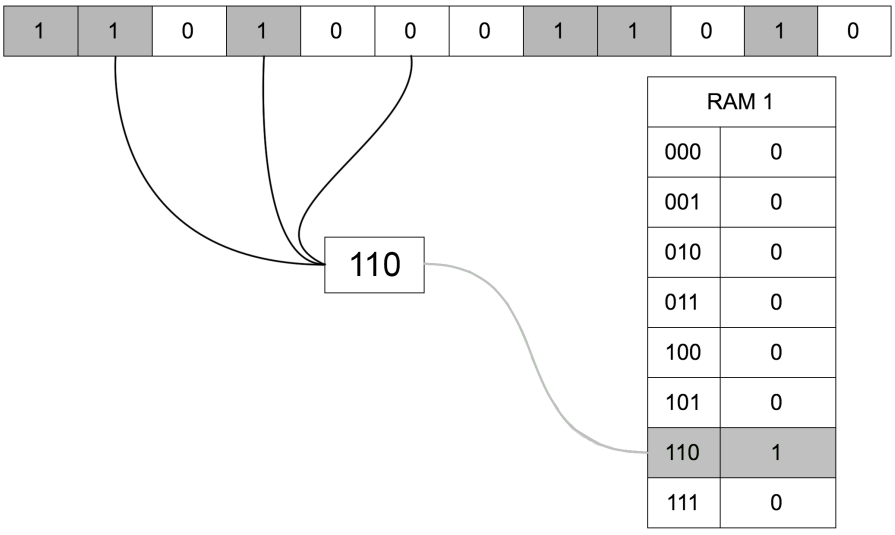


Figura 4. Funcionamento de uma *RAM*

A *RAM* faz isso, pois está reconhecendo que a classe de seu discriminador, possui os valores vistos na retina nas suas respectivas posições (ou seja, cada *RAM* fica responsável por um sub-padrão). Assim, quando a rede estiver em modo de classificação, a *RAM* olhará para a sua parte da retina e verá se o endereço formado pelos valores vistos possui o valor 1. Se o valor no endereço for 1, significa que a *RAM* viu aquele sub-padrão durante o treinamento, e portanto, em sua visão limitada da retina, a entrada é da classe de seu discriminador. Se o valor do endereço for 0, significa que aquele sub-padrão não

foi visto durante o treinamento e para a *RAM* a entrada não faz parte da classe de seu discriminador.

Como um todo, os discriminadores possuem *RAMs* que votam dizendo se a entrada pertence ou não à sua classe. No modo de classificação, as *RAMs* dizem que a entrada pertence à classe de seu discriminador passando um 1 ao discriminador, e dizem que não pertence passando um 0 ao discriminador. Os discriminadores somam todos os seus votos e aquele que tiver o maior valor é considerado a classe da entrada. Ou seja, o discriminador que tiver mais *RAMs* dizendo que a entrada pertence à sua classe é escolhido como resposta certa [4][5]. Com isso, é possível ter uma classificação das entradas dadas as classes possíveis. Mais do que isso, é possível extrair uma confiança da classificação feita pela rede. Basta compararmos as pontuações de seus discriminadores. Se o resultado for muito próximo entre classes diferentes, significa que a rede não soube diferenciar bem as classes para a entrada apresentada. Uma fórmula que determina a confiança da *WiSARD* é dada por (3),

$$C = \frac{score(Classe1) - score(Classe2)}{score(Classe1)} \tag{3}$$

onde C é a confiança da classificação, $score(Classe1)$ a pontuação da classe com maior pontuação (escolhida como resposta) e $score(Classe2)$ a pontuação da classe com a segunda maior pontuação. Na Figura 5, d representa a diferença entre as duas maiores pontuações.

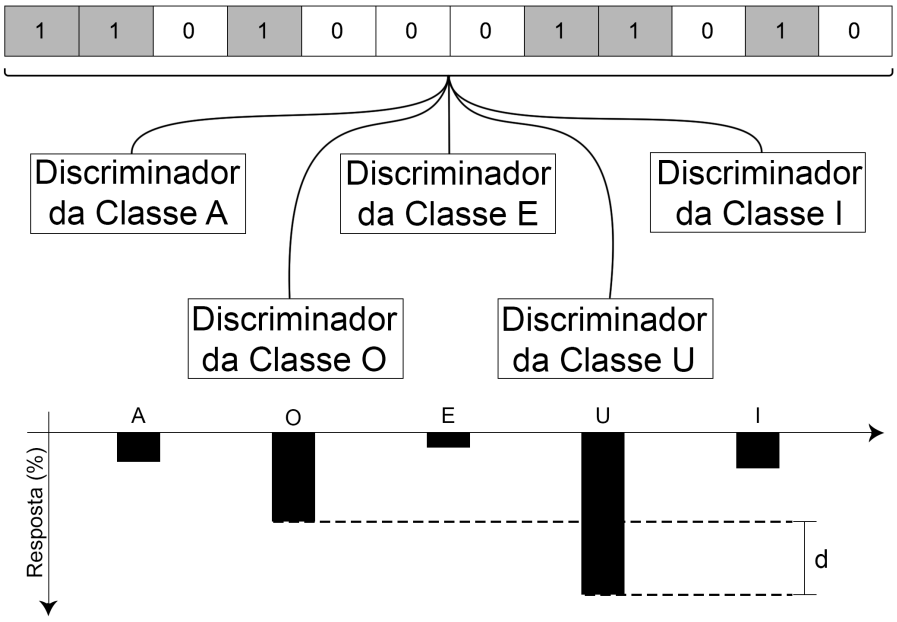


Figura 5. Pontuação dos discriminadores de cada classe na *WiSARD*

Há uma variação da rede chamada *WiSARD* com *bleaching*. Nessa variação, em vez de cada *RAM* apenas marcar 1 nas posições de memória vistas durante o treinamento, cada *RAM* incrementa 1 no valor prévio da posição de memória. Isso corrige um problema presente na *WiSARD* original. O problema é que se mostrarmos muitos exemplos de uma mesma classe durante o treinamento, cada um com leves mudanças em relação aos outros,

a *RAM* pode ter muitas posições marcadas com 1, ou seja, a *RAM* votará 1 durante a classificação para praticamente qualquer entrada que receber, uma vez que já viu de tudo durante o treinamento. Com a técnica de *bleaching*, as *RAMs* só votam 1 no momento da classificação se o valor lido da posição de memória construída a partir da entrada for maior que um limiar. Esse limiar começa baixo, normalmente em zero. Se a confiança da classificação for baixa, aumenta-se o limiar e tenta-se a classificação novamente. Esse processo se repete até que a confiança esteja acima de um valor estipulado, ou que o máximo valor salvo nas RAMs seja alcançado. Neste caso, a resposta é aquela que obteve a maior confiança, ainda que não tivesse sido a ideal.

3. Conclusão

Atualmente a biblioteca se encontra próxima de ser concluída, foi arquitetada e desenvolvida seguindo os preceitos encontrados em [1] e [2]. Após a conclusão da mesma serão feitos experimentos de reconhecimentos de elocuções com gravações de mais indivíduos, tanto daqueles hábeis, quanto daqueles com necessidades especiais, executando assim a validação do modelo. Assim que estiver finalizada as etapas seguintes serão de iniciar os experimentos mais detalhados que serão reportados em comunicações científicas na fase final da pesquisa. Na etapa em que o projeto se encontra, a gravação das elocuções é feita de maneira externa ao projeto, com os áudios sendo gravados em arquivos no formato *WAV* que serão processados pela biblioteca. Existem planos futuros para integrar a funcionalidade de gravação e processamento de um fluxo de áudio em tempo real na biblioteca.

Referências

[1] D. F. P. DE SOUZA, F. M. G. FRANCA, and P. M. V. LIMA. Spatio-temporal pattern classification with kernelcanvas and wisard. *Brazilian Conference on Intelligent Systems*, pages 228–233, October 2014.

[2] D. F. P. DE SOUZA. Time-series classification with kernelcanvas and wisard. master thesis, UFRJ/COPPE, Rio de Janeiro – RJ – Brasil, 2015.

[3] J. Lyons. Mel frequency cepstral coefficient (mfcc) tutorial, 2013. `http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/` Acessado em: 05/09/2017.

[4] I. ALEKSANDER, M. DE GREGORIO, F. M. G. FRANCA, P. M. V. LIMA, and H. MORTON. A brief introduction to weightless neural systems. In *European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning*, volume ESANN'2009 proceedings, Bruges (Belgium), 22-24 April 2009. ESANN. d-side publi.

[5] I. ALEKSANDER, W. V. THOMAS, and P. A. BOWDEN. Wisard: A radical step foward in image recognition. *Sensor Review*, 4:120–124, July 1984.

[6] F. J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. In *Proceedings of the IEEE*, volume Volume: 66, Naval Ocean Systems Center, San Diego, CA, Jan 1978. IEEE.

Explorando Computação Evolutiva em *Workflows* Científicos

Anderson Ferreira[1], Fabricio Firmino Farias[2], Sérgio Manuel Serra da Cruz[1,3]

[1] Programa de Pós-Graduação em Matemática Computacional - PPGMMC/UFRRJ [1]
[2]Departamento de Ciência da Computação – UFRJ
[3] Departamento de Computação – UFRRJ

anderson@infoassociados.com.br, firminodefaria@ufrj.br, serra@pet-si.ufrrj.br

Resumo. *A computação evolutiva está se tornando uma alternativa aos modelos computacionais clássicos, uma vez que é capaz de resolver eficientemente problemas no tempo polinomial. Essa área é fortemente baseada no comportamento biológico de organismos vivos, usando elementos da natureza como uma base de computação para obter melhores soluções para resolver problemas complexos. Por outro lado, os tradicionais workflows científicos vêm ganhando espaço na computação, entretanto, ainda oferecem suporte de forma limitada para a computação evolutiva. Para preencher esta lacuna, este artigo apresenta os primeiros estudos da abordagem VisPyGMO que incorpora os algoritmos evolutivos em SGWfCs usando o paradigma genérico de modelo de ilha. Ao término do estudo apresentamos um caso de uso real usando tais algoritmos no sistema VisTrails.*

Abstract. *Evolutionary computing is becoming an alternative to classical computational models since its able to solve, in an efficient way, hard problems in polynomial time. This area is based on the biological behavior of living organisms, using nature as a basis of computation to obtain better solutions to solve complex problems. On the other hand, the traditional scientific workflows systems offer little support to evolutionary computing. To bridge this gap, this paper describes the VisPyGMO approach which extends the existing workflow systems by incorporating evolutionary algorithms using the generic island model paradigm. Besides, we present a real use case using such algorithms in VisTrails system.*

1. Introdução

Os avanços da ciência da computação possibilitaram que os pesquisadores de diversas áreas do conhecimento utilizem simulações computacionais em suas pesquisas, obtendo resultados de qualidade em menor tempo e com menores custos. Muitas dessas simulações envolvem diversos tipos de paradigmas da computação que podem variar desde algoritmos clássicos baseados em modelos matemáticos simples até complexos modelos de otimização multiobjetivos apoiados por Computação Evolutiva (CE) (Back *et al*., 1997).

Muitas simulações computacionais são modeladas por *workflows* científicos. Um *workflow* se caracteriza por uma abstração capaz de representar experimentos científicos, onde um pesquisador define quais programas serão executados, a ordem e os parâmetros que os programas utilizarão e suas dependências de dados (Deelman et al., 2009). Os *workflows* são gerenciados por Sistemas de Gerência de *Workflows* Científicos (SGWfC). Atualmente, existem diversos SGWfC, Esses sistemas são responsáveis por diversas funções, a saber: gerenciar a execução, distribuir sua

execução em ambientes de processamento de alto desempenho, coletar dados de proveniência (Freire *et al*. 2008).

Algoritmos baseados em Computação Evolutiva são uma família de métodos computacionais que buscam soluções de problemas do mundo real baseando-se nas soluções que a própria natureza encontrou durante o curso do processo evolutivo. As abordagens da CE são variadas, propõe um paradigma inspirado na teoria de seleção natural proposta por Darwin (Back *et al.,* 1997), métodos baseados em inteligência de enxames, como *Particle Swarm Optimization* (PSO) (Kennedy e Eberhart, 1995), *Fish School Search* (FSS), *Firefly Algorithm* (FA), *Ant Colony* (AC) e *Artificial Bee Colony* (ABC) (Karaboga, 2005) e métodos baseados em processos físicos, como o *Simulated Annealing* (SA) (Kirkpatrick *et al*., 1983). Resumidamente, os métodos têm como objetivo encontrar boas soluções, sem garantia de solução ótima, para problemas que envolvem muitas variáveis, com múltiplos objetivos e que apresentam diversas restrições, características comuns para muitos tópicos de pesquisa atuais nas mais variadas áreas do conhecimento.

Apesar de os SGWfCs oferecerem um arcabouço refinado para especificação, execução e monitoramento de *workflows* em ambientes centralizados ou distribuídos, estes ainda carecem de suporte aos algoritmos bioinsirados (ABI), principalmente no que tange ao apoio à composição dos experimentos baseadas em CE. Dessa forma, se torna crucial a busca e oferta de novas soluções para essa limitação.

Diversas áreas da otimização utilizam algoritmos evolucionários que emergiram de experiências baseadas CE que modelam o "comportamento social" de muitas espécies de insetos, pássaros e peixes (Kennedy e Eberhart, 1995). A busca de soluções quase-ótimas baseadas em algoritmos evolucionários é um problema recorrente em diversas áreas: Engenharia, Economia, Química e *Big Data* (Cao *et al.,* 2016). Assim, apoiar a construção de *workflows* que utilizam algoritmos evolutivos para a solução de problemas de otimização multiobjetivos é uma tarefa de grande importância. Atualmente, diversas bibliotecas são oferecidas. No entanto, por serem desconectadas dos SGWfC elas ainda não se beneficiam das facilidades oferecidas por esses sistemas, por exemplo, (re)xecução parcial, coleta de dados proveniência, monitoramento de execução, entre outros.

Nesse artigo, o objetivo é apresentar uma nova abordagem denominada VisPyGMO que simplifica a composição de *workflows* baseados em algoritmos de CE. VisPyGMO visa auxiliar pesquisadores a modelar *workflows* científicos que tratam de complexos problemas de otimização. A ideia principal por trás da abordagem é incorporar a um SGWfC clássico um conjunto de componentes genéricos que reutilizam algoritmos evolucionários, permitindo que pesquisadores da área de inteligência computacional e otimização se beneficiem das facilidades oferecidas pelos SGWfCs. Todo o arcabouço proposto foi implementado no SGWfC Vistrails (Callahan *et al*., 2006).

Esse artigo está organizado da seguinte forma. A seção 2 apresenta o referencial teórico. A seção 3 apresenta a abordagem VisPyGMO. A seção 4 contém as avaliações experimentais de um estudo de caso baseado em *datasets* reais extraídos da plataforma de desafios de modelagem cognitiva e aprendizado de máquina *Kaggle* (Monajemi et al., 2016). A seção 5 discute os trabalhos relacionados. Finalmente, a seção 6 apresenta as conclusões destacando as contribuições e trabalhos futuros.

2. Fundamentação Teórica

A CE tem sido empregada em uma variedade de disciplinas, desde ciências naturais e engenharia até biologia e ciência da computação. A CE compreende um conjunto de técnicas de busca e otimização inspiradas na evolução natural das espécies. Atualmente as principais técnicas incluem os algoritmos genéticos (Holland, 1992) e a inteligência de exames (Blum e Maerkle 2008 e Panigrahi *et al*., 2011).

2.1. Inteligência de Exames

Segundo Blum e Maerkle (2008), existem várias fontes biológicas da inteligência, por exemplo, colônias de formigas, enxames de abelhas, cardumes de peixes e bandos de pássaros, entre outros. Neste artigo adotamos a definição clássica de Kennedy e Eberhsrt (1995), se considera que a inteligência de enxames é uma tentativa de criar algoritmos para solução de problemas distribuídos (inspirados no comportamento coletivo de populações de insetos sociais ou sociedades de animais) que sejam de baixo custo, rápidos e ofereçam soluções robustas para problemas complexos.

Na literatura existem diversos algoritmos de otimização baseados em inteligência de enxames. Conceitualmente eles podem ser representados como populações conhecidas submetidas a métodos de exploração orientada (Luque e Alba, 2011). Uma população inicial de soluções candidatas é aprimorada por meio de variações realizadas em seus indivíduos e pelo processo de seleção (Figura 1). Os algoritmos mais disseminados são: PSO, ABC, SA e *Differential Evolution* (DE) (Storn e Price, 1997).

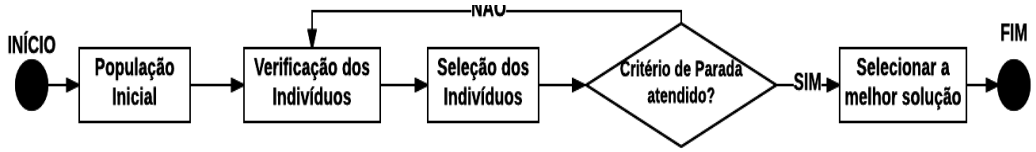


Figura 1. Diagrama conceitual do fluxo comum aos algoritmos de CE.

Os algoritmos PSO são de natureza estocástica e consideram os membros de uma comunidade como uma partícula. Computacionalmente, fazendo uma analogia, o termo partícula foi adotado para simbolizar os pássaros e representar as possíveis soluções do problema a ser resolvido. A área sobrevoada pelos pássaros é equivalente ao espaço de busca e encontrar o local com comida, ou o ninho, corresponde a encontrar a solução ótima. Dentre as aplicações do PSO, se destacam os modelos híbridos para auxiliar a melhoria de desempenho de outros métodos e algoritmos de otimização. Outra linha é a utilização do PSO na resolução de problemas de uma ampla gama de aplicações, tais como: sintonização de controladores, otimização de controle de potência reativa e tensão, despacho econômico de carga, detecção de falhas, projeto de antenas, jogos, processamento de imagens, redes de sensores e mineração de dados (Serapião, 2009).

Os algoritmos ABC são inspirados nas colônias de abelhas melíferas durante o ato de forrageamento, onde as abelhas possuem três tipos de comportamento. As abelhas trabalhadoras (exploram as fontes de néctar e compartilham informações ao retornarem para a colmeia); as exploradoras (realizam buscas randômicas para encontrar novas fontes) e as oportunistas (aguardam na colmeia, escolhem uma fonte para explorar de acordo com as informações recebidas das abelhas exploradoras). Através da interação entre oportunistas e trabalhadoras desenvolvem uma inteligência coletiva que otimiza a sua busca de alimentos. Computacionalmente, a posição de uma fonte de néctar é representada por uma solução do espaço de busca para o problema e a quantidade de néctar dessa fonte representa o valor da função aptidão dessa solução. Esse algoritmo é comumente empregado em problemas de otimização com espaço de busca contínuo.

Os algoritmos SA estão baseados em processos termodinâmicos; simulam o processo de recozimento de metais onde a variação de temperatura pode conduzir a metais mais estáveis, estruturalmente fortes e de menor energia. Computacionalmente, há uma analogia com um problema combinatório, onde os possíveis estados de um metal correspondem a soluções do espaço de busca. A energia em cada estado corresponde ao valor da função objetivo. A energia mínima (ou máxima) corresponde ao valor de uma solução ótima local, possivelmente global. Entre as aplicações do SA, se destacam aquelas voltadas à resolução de problemas de otimização combinatória.

Por fim, os algoritmos DE são de natureza estocástica. Eles têm como objetivo resolver o problema de ajuste polinomial de Chebyshev. Computacionalmente, durante as iterações, o DE altera as coordenadas das soluções (vetores) candidatas a partir de um vetor gerado pela adição de um vetor à diferença ponderada entre outros dois vetores. Esse algoritmo é comumente empregado em problemas de computação paralela e de otimização uni ou multiobjetivos.

2.2. Modelos em Ilhas

O Modelo em Ilhas (IM), descrito em Izzo *et al*., (2012) é um paradigma multipopulacional para melhorar o desempenho dos algoritmos evolucionários na resolução de problemas de otimização. O modelo faz a distribuição eficiente dos algoritmos em múltiplos processadores e pode ser utilizado em arquiteturas paralelas ou não. Sua inspiração se baseia na teoria do Equilíbrio Pontuado (Cohoon *et al*., 1987), que tem o intuito de explicar certos dilemas paleontológicos ao longo do registro geológico.

No IM, periodicamente ocorre uma troca de indivíduos entre as ilhas vizinhas de um arquipélago (migração). O processo de migração cria fluxos direcionados entre as ilhas do modelo, denominadas topologias de migração. Estes fluxos migratórios formam a estrutura de vizinhança do IM. As topologias de migração representam uma importante regra do IM e podem exercer grande influência no seu desempenho (Izzo *et al*., 2012).

Assim, em projetos de IM, a população total é dividida em subpopulações em ilhas (Luque e Alba, 2011) e cada ilha executa um algoritmo independentemente e mantém sua própria subpopulação para pesquisa orientada.

3. Abordagem Proposta

Nessa seção apresentamos a abordagem proposta, denominada VisPyGMO. A abordagem consiste oferecer um conjunto de módulos que se baseiam nos algoritmos de CE (apresentados na seção 2) que podem ser acoplados a um SGWfC tradicional. Em sua versão atual o VisPyGMO foi configurado para ser acoplado ao SGWfC VisTrails.

O VisPyGMO utiliza o modelo em Ilha de Izzo *et al*. (2012) que pode ser aplicado a uma família de problemas de otimização para encapsular os algoritmos evolucionários. Os algoritmos utilizados nessa implementação estão disponíveis nas bibliotecas PaGMO/PyGMO de código aberto e com suporte à computação paralela (Izzo *et al*., 2012 e PyGMO, 2017).

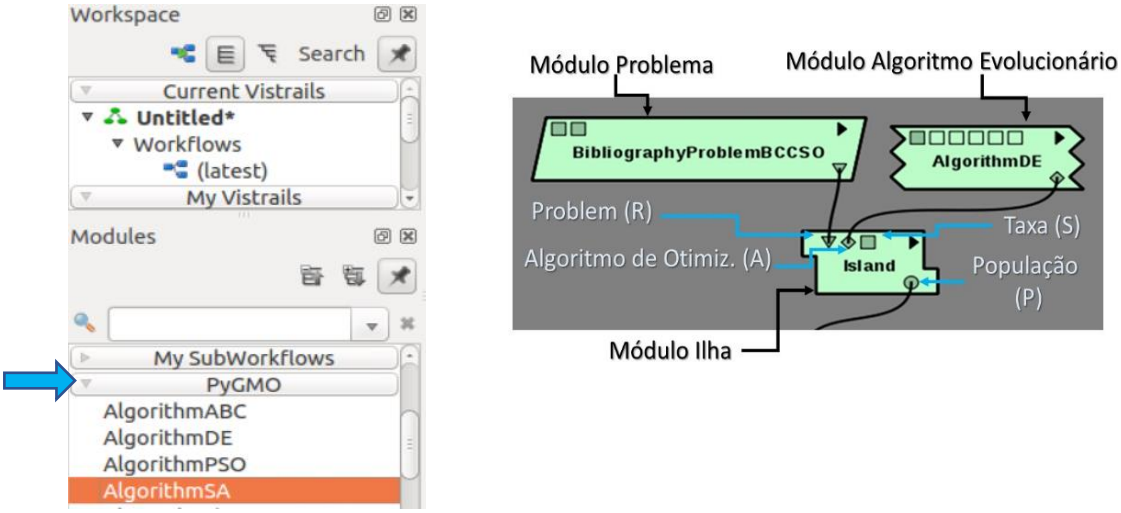


Figura 2- Pacote PyGMO no SGWfC VisTrails (à esquerda), Fragmento de *workflow* com módulos AlgorithmDE, Ilha e a indicação das portas (à direita).

Optou-se pelo reuso dessas bibliotecas pelos seguintes motivos. (i) Oferecem mais de 20 tipos de algoritmos de otimização de propósito geral distribuídos em três categorias (heurísticos, meta-heurísticos e de otimização local) que podem ser encapsulados no modelo em ilha; (ii) Possuem suporte nativo ao *Message Passing Interface* (MPI). Por exemplo, em um processo de otimização se podem abrir múltiplas *threads* e explorar o paralelismo da máquina local ou fazer distribuição de carga através de chamadas MPI em ambientes de processamento paralelo (*clusters* ou nuvens de computadores). No caso, ambos os motivos se alinham com as características desejáveis de utilização de *workflows* científicos na E-Ciência (Deelman *et al*., 2009).

A primeira versão do VisPyGMO foi criado no SGWfC VisTrails um pacote denominado PyGMO composto por quatro módulos principais, cada um representando um algoritmo evolucionário (PSO, SA, ABC e DE) (Figura 2 à esquerda), também foi implementado o módulo genérico do modelo em ilha denominado *Island*. Ele é composto por 4 portas lógicas, cada uma corresponde a um dos parâmetros da equação (2). Todos os pacotes e módulos foram implementados em linguagem Python e os algoritmos apresentados na Seção 2 em C++ e Python.

4. Estudo de Caso

Como prova de conceito do VisPyGMO foram realizados vários experimentos. A experimentação utilizou openSUSE versão 42.2, biblioteca PaGMO/PyGMO 1.1.5 e Python 2.7 do VisTrails 2.2.4. O hardware utilizado foi um processador i5 com 12GB RAM e 128GB HD SSD.

Os experimentos tiveram como objetivo avaliar a funcionalidade dos módulos CE ao tentar responder desafios de *data analytics* disponibilizados pela plataforma *Kaggle*. Em linhas gerais tais desafios são problemas reais postados por empresas ou governos e que demandam abordagens de gerenciamento de grandes volumes de dados e uso de técnicas de CE ou de aprendizado de máquina. Especificamente, o desafio proposto está relacionado com o *Pesticide Data Program* (PDP) dos EUA, onde o Departamento de Agricultura (PDP, 2017) analisa resíduos de pesticidas encontrados em alimentos (detalhes em https://goo.gl/WtBZZj).

O desafio tratado neste artigo pode ser resumido da seguinte forma: “Quais os pesticidas mais comumente utilizados nos EUA e quais as futuras tendências de consumo destes nos próximos 5 anos?”

Para que os experimentos de CE fossem executados, foi necessário consolidar vários *datasets* do PDP. Originalmente, os dados brutos se encontravam dispersos em longas séries (desde 1995 até 2015), totalizando uma população de mais de 40 milhões de análises de resíduos e 670 de tipos de pesticidas em aproximadamente 80 MB de arquivos CSV.

Tabela 1- Ranking dos tipos de pesticidas utilizados entre 1995-2015 nos EUA.

Rank	Nome	∑ nº análises de resíduos	Rank	Nome	∑ nº análises de resíduos
1	*Malathion*	233.044	**6**	*Diazinon*	210.406
2	*Chlorpyrifos*	227.504	**7**	*Dimethoate*	201.552
3	*Carbaryl*	224.723	**8**	*Endosulfan I*	201.483
4	*Methomyl*	211.307	**9**	*Endosulfan II*	201.184
5	*Myclobutanil*	211.294	**10**	*Endosulfan sulfate*	200.236

Na primeira parte do desafio, desenvolvemos um *workflow* genérico (não ilustrado no texto) que carrega as séries, trata os dados e gera todas as ilhas, ou seja, grupamentos de indivíduos de uma mesma categoria de pesticidas. Quantitativamente, os resultados são apresentados na Tabela 1. Por limitações de espaço, nesta subseção discutiremos apenas os estudos de tendências sobre a categoria (ilha) mais populosa (*Malathion)* composta por 6.954 indivíduos selecionados pelo *workflow*.

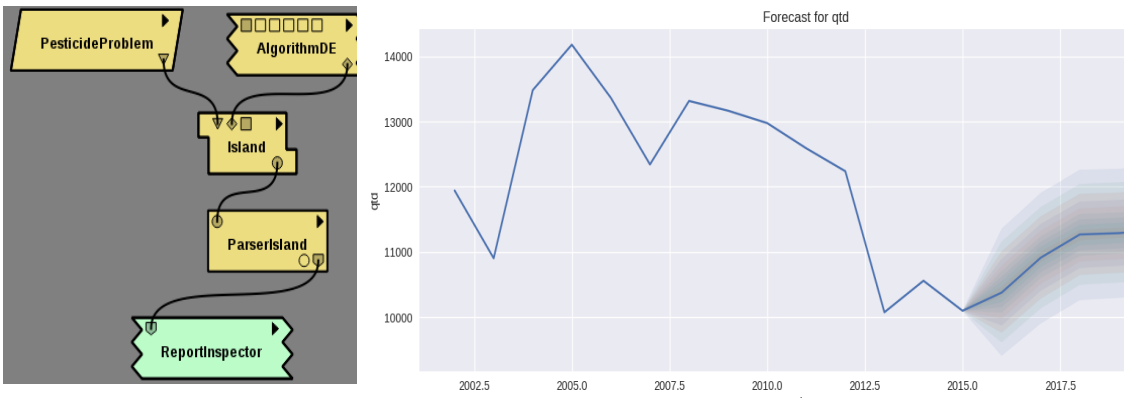


Figura 3. – Fragmento do *workflow* de predição de uso de pesticidas (à esquerda), visualização das projeções do *Malathion* no quinquênio (à direita).

Na segunda parte do desafio desenvolvemos um outro *workflow* genérico composto de módulos PyGMO. Por se tratar de um problema de ajuste dos hiperparâmetros do preditor, o *workflow* requer o uso dos módulos *Differential Evolution* (DE) e modelo Ilha (Figura 3 à esquerda). O módulo *Pesticide Problem* modela a carga das ilhas. O *Parse Island* interpreta os dados populacionais de cada ilha. O *Report Inspector* gera os relatórios com valores de preditos das quantidades dos pesticidas em séries temporais. Adotou-se uma projeção gráfica do tipo *Autoregressive Integrated Moving Average* (ARIMA), da biblioteca PaGMO/PyGMO, para a visualização das séries de consumo dos pesticidas para o período de 2016-2020. Os experimentos obtiveram uma taxa de erro e margem de tolerância de 95% no cálculo das projeções.

5. Trabalhos Relacionados

Existem algumas abordagens na literatura que correlacionam CE com SGWfC. No entanto, nenhuma delas tem como objetivo facilitar a composição de experimentos com algoritmos evolucionários apoiados por *workflows* científicos. Dessa forma, apresentamos nesta seção as abordagens que utilizam princípios de CE.

Em geral a CE é muito utilizada pelos escalonadores dos SGWfC em ambientes distribuídos. Por exemplo, o Askalon é um SGWfC cujo objetivo principal é simplificar a execução de *workflow* em grades computacionais (Wieczorek *et al.,* 2005). Ele implementa um conjunto de algoritmos de escalonamento que difere dos outros SGWfCs por sua diversidade de tipos. Por exemplo, o escalonador de tarefas do Askalon provê tanto algoritmos com desempenhos garantidos quanto heurísticas baseadas em algoritmos genéticos. Recentemente, Gao *et al.* (2010) e Zhu *et al.* (2016) investigaram o problema do escalonamento em ambiente de nuvem e propuseram a utilização de algoritmos de otimização bioinpirados como alternativa de escalonamento em nuvens do tipo Infraestrutura como serviço (IaaS).

No caso específico do VisTrails, verificou-se que o sistema oferece módulos de algoritmos de aprendizado de máquina através da biblioteca *scikit-Learn*, porém não se verificou suporte equivalente para CE. Ao investigarmos outros sistemas e SGWfC clássicos que apoiam a experimentação científica (*e.g.* CodaLab, SDM, Kurator, IH, Torch, Sumatra, Kepler, Pegasus) também se verificou que eles não oferecem suporte a componentes de CE.

6. Considerações Finais

Novas possibilidades para as pesquisas se abrem com difusão do uso de *workflows* científicos como abstração para modelar experimentos baseados em problemas de otimização que demandam algoritmos evolucionários. Esse artigo apresentou uma contribuição para a área de E-Ciência baseada em elementos tradicionais da área da CE.

O VisPyGMO permite que pesquisadores utilizem algoritmos evolucionários em SGWfC de modo simplificado. Para isso, como prova de conceito reutilizamos a biblioteca PaGMO/PyGMO, incorporamos módulos CE baseados no Modelo em Ilha no SGWfC Vistrails e aplicamos o conceito na resolução de um desafio proposto pela plataforma *Kaggle*. Como trabalhos futuros se pretende incluir novos algoritmos de otimização ao VisTrails, por exemplo, algoritmos genéticos, buscas de Monte Carlo, entre outros. Além disso, se efetuarão sobre proveniência de dados (Cruz, Campos e Mattoso, 2009) e avaliações de desempenho dos módulos já implementados em *workflows* sendo executados em ambientes paralelos (Cardozo, Thomaz e Cruz, 2016).

Agradecimentos

Os autores agradecem ao programa ao Programa de Educação Tutorial, ao FNDE, a Red *BigDSSAgro CYTED* e *Microsoft Azure Research* (CRM:0518152) pelas bolsas e financiamentos concedidos.

Referências

Bäck, T., Fogel, D.B.; Michalewicz, Z. (1997) Handbook of Evolutionary Computation, Institute of Physics Publishing and Oxford University Press.

Blum C., Merkle D. (2008). Swarm Intelligence – Introduction and Applications. Natural Computing. Springer, Berlin.

Cao, J., *et al*. (2016) Big Data: A Parallel Particle Swarm Optimization-Back-Propagation Neural Network Algorithm Based on MapReduce. PLoS One 11(6): e0157551.

Callahan, S. P. *et al*. (2006). VisTrails: visualization meets data management. In: Proc. SIGMOD 2006, pp. 745-747, USA.

Cardozo, F.; Tomaz, U. R.; Cruz, S. M. S. (2016) Avaliando Uma Estratégia Computacional Baseada Em Workflows Científicos Apoiados Por Placas Gráficas Genéricas. In: ERSI 2016 p. 66-73, Brasil.

Cohoon, J. P., *et al*. (1987). Punctuated equilibria: a parallel genetic algorithm, Proc. of the Int. Conf. on Genetic Algorithms and their application, NJ, USA, pp. 148–154.

Cruz, S.M.S. *et al.* (2009). Towards a Taxonomy of Provenance in Scientific *Workflow* Management Systems, In: Congress on Services, IEEE, pp. 259–266.

Deelman, E., Gannon, D. Shields, M., Taylor, I. (2009). Workflows and e-Science: An overview of workflow system features and capabilities. FGCS, vol. 25, no. 5, pp. 528–540.

Freire, J., Koop, D., Santos, E., Silva, C. T. (2008). Provenance for Computational Tasks: A Survey, Computing in Science and Engineering, vol. 10, pp. 11–21.

Gao, Y., Hu, X., Liu, H., Li, F. (2010). Cloud Estimation of Distribution Particle Swarm Optimizer, 4th Int. Conf. on Genetic and Evolutionary Computing, pp.1-14.

Holland, J.H. (1992). Adaptation in Natural and Artificial Systems. 2nd edition, MIT Press.

Izzo, D., Ruciński, M. and Biscani, F. (2012). The Generalized Island Model. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 151–169.

Karaboga, D. (2005). An Idea Based On Honey Bee Swarm For Numerical Optimization. Tec. Report TR06.

Kennedy J., R. C. Eberhart. (1995). Particle Swarm Optimization. In Proc. of IEEE International Conference on Neural Networks, Perth, Australia, pp. 1942–1948.

Kirkpatrick, S., Gelatt C. D., Vecchi M. P., (1983). Optimization by Simulated Annealing, Science, Vol 220, Number 4598, pp. 671-680.

Luque, G., Alba, E. (2011). Parallel Genetic Algorithms: Theory and Real World Applications, Studies in Computational Intelligence, Springer.

Monajemi H.; Donoho D. L.; Stodden V. (2016). Making Massive Computational Experiments Painless. Int Conf Big Data, 5pp.

Panigrahi, B. K. Shi Y., Lim M.-H. (2011): Handbook of Swarm Intelligence. Series: Adaptation, Learning, and Optimization, vol. 7, Springer-Verlag Berlin.

PDP (2017). https://www.ams.usda.gov/datasets/pdp/pdpdata. Acesso em 05 jan. 2017.

PyGMO, (2017). http://pagmo.sourceforge.net/pygmo/documentation/. Acesso em 10 fev. 2017.

Serapião, A. B. S (2009). Fundamentos de otimização por inteligência de enxames: uma visão geral. Controle & Automação. vol. 20 no.3.

Storn, R.; Price, K. (1997). Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optimization. 11: pp. 341–359.

Wieczorek, M., Prodan, R., Fahringer, T. (2005). Scheduling of Scientific Workflows in the ASKALON Grid Environment. SIGMOD Rec., vol. 34, pp. 56-62.

Zhu, Z., Zhang, G., Li, M., Liu X. (2016). Evolutionary Multi-Objective Workflow Scheduling in Cloud. IEEE Transactions on Parallel and Distributed Systems, vol. 27, 5.

Uma investigação sobre estratégias a serem adotadas para o aprendizado de Inteligência Artificial no Ensino Fundamental por meio da Robótica Educacional

Rubens L. Queiroz[1], Fábio F. Sampaio[1,2], Priscila M. V. Lima [1,2]

[1]Programa de Pós-Graduação em Informática (PPGI/UFRJ)

[2] Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais (NCE/UFRJ)

rubensqueiroz@outlook.com, ffs@nce.ufrj.br, priscila.lima@nce.ufrj.br

Abstract. *This paper presents an initial bibliographic review about the investigation of strategies to be adopted for the learning of Artificial Intelligence (AI) in Elementary School, because it is considered a topic that assumes significant relevance in the current context, in which more and more jobs will have some kind of relationship with AI. It is presented here a synthesis of eleven works around this theme, described some possibilities of approaches to be adopted in relation to AI learning in Elementary School and presented some suggestions for future work within this scope.*

Resumo. *Este trabalho apresenta uma revisão bibliográfica inicial acerca da investigação de estratégias a serem adotadas para o aprendizado de Inteligência Artificial (IA) no Ensino Fundamental, por considerar-se esse um tema que assume significativa relevância no contexto atual, no qual cada vez mais postos de trabalho passarão a ter algum tipo de relação com IA. É apresentada aqui uma síntese de onze trabalhos em torno deste tema, descritas algumas possibilidades de abordagens a serem adotadas em relação ao aprendizado de IA no Ensino Fundamental e apontadas sugestões de trabalhos futuros dentro deste escopo.*

1. Introdução

A Inteligência Artificial (IA) desempenha atualmente um importante papel na vida cotidiana, estando presente nos mais diversos setores. Nos próximos anos, cada vez mais postos de trabalho passarão a ter algum tipo de relação com IA. Esta realidade faz com que o desenvolvimento de conhecimentos neste campo passe a ter grande importância para profissionais de diversas áreas. Neste contexto, a "alfabetização" em IA assume fundamental importância [Kandlhofer *et al*. 2016].

> A alfabetização clássica permite que as pessoas leiam e compreendam novos textos, ao invés de simplesmente decorá-los [Sklar e Parsons 2002]. O mesmo se aplica à alfabetização em IA: ela permite que as pessoas compreendam técnicas e conceitos existentes por trás dos produtos e serviços de IA ao invés de apenas aprenderem como utilizar determinadas tecnologias ou novas aplicações [Kandlhofer *et al*. 2016, p.2, tradução nossa].

De acordo com os dados levantados por Kandlhofer *et al.* (2016), são muito raros os estudos realizados sobre o ensino de IA na Educação Básica. Os autores afirmam ainda que pesquisas na área de alfabetização mostram que, para os alunos serem capazes de desenvolver habilidades avançadas em relação à leitura e escrita, é fundamental iniciar o

trabalho em relação à essas habilidades desde a tenra idade, e que o mesmo, acredita-se, seja válido no que se refere ao aprendizado de IA

Esse quadro aponta para a necessidade do desenvolvimento de ferramentas que possam dar apoio a essa "alfabetização em IA", inclusive como parte integrante da Educação Básica, no que, o emprego da Robótica Educacional, possivelmente, seja uma estratégia bastante interessante de ser adotada, uma vez que seu uso vem apresentando resultados muito promissores no que ser refere ao aprendizado de programação de computadores e de outros temas relacionados tanto às ciências da computação quanto a outras áreas do conhecimento, devido, entre outras questões, aos seus caráteres Construtivista e Construcionista, ou seja, de desenvolvimento do aprendizado a partir da experimentação e da criação [Queiroz 2017]. A robótica é, na verdade, uma faceta fundamental no "empreendimento" da IA [Dodds *et al*. 2006], o que já demonstra a importância do uso da Robótica Educacional em cursos destinados ao aprendizado desta disciplina.

Em conjunto com o desenvolvimento de ferramentas, faz-se primordial a realização de pesquisas que procurem auxiliar na busca por respostas à questionamentos tais como: Que conceitos relacionados à IA podem ser trabalhados no Ensino Fundamental por meio da Robótica Educacional? Em que profundidade esses conceitos devem ser trabalhados? Dentre eles, quais são adequados de serem abordados a partir de qual idade? É com o intuito de coletar informações preliminares que possam apontar caminhos a serem seguidos na busca por respostas a estes e outros questionamentos, relacionados à temática do aprendizado de IA no Ensino Fundamental por meio da Robótica Educacional, que se concentra o objetivo do estudo aqui apresentado.

Este trabalho está organizado da seguinte forma: na próxima seção é descrito o método adotado para o levantamento dos dados, na sessão 3 é feita uma síntese dos dados coletados, na seção 4 são apresentadas algumas conclusões acerca desses dados e, por fim, na última seção são apontadas algumas possibilidades de pesquisas futuras.

2. Método

O método de pesquisa adotado por este estudo foi a realização de uma revisão bibliográfica acerca dos temas investigados. Para tal, efetuou-se uma busca nas bases Periódicos Capes e Google Acadêmico a partir de um conjunto de *strings* de busca construídas com base na seguinte combinação de palavras-chave: (*artificial intelligence* OR *computational intelligence* OR *machine learning* OR *intelligent agents* OR *agents*) AND (*educational robotics* OR *beginners* OR *dummies* OR *arts* OR *makers* OR *maker universe* OR *children* OR *k-12*[1] OR *primary school*).

Os artigos selecionados nesta busca, com base na leitura dos títulos e resumos dos trabalhos mais relevantes[2] retornados, foram então organizados em uma planilha, ordenados incialmente por *string* de busca e, posteriormente, aglutinados por "tema" (ver Quadro 1). Estes temas foram escolhidos, dentre aqueles abordados nos artigos selecionados, por acreditar-se que as pesquisas em torno deles poderiam apontar possíveis

[1] A sigla K-12 é utilizada, nos Estados Unidos e em alguns outros países de língua inglesa, para designar o Ensino Fundamental e Médio como um todo

[2] Relevância indicada pelos motores de Busca.

recursos e estratégias pedagógicas a serem adotados para o aprendizado de IA, mais especificamente por meio da Robótica Educacional, no Ensino Fundamental.

Para cada artigo incluído na planilha, foram inseridas as seguintes informações: ordem em que foi "encontrado", *string* de busca, motor de busca (Capes /Google), ano, autor, resumo, tema, observações, link para o artigo e sigla (um código identificador par o documento). Por meio desse processo foram selecionados 19 artigos, agrupados em 4 temas, conforme apresentado no Quadro 1.

Finalizada essa fase, foram então realizadas as leituras das introduções e conclusões dos 19 artigos e, para fins dessa investigação inicial, selecionados 9 trabalhos para a realização de uma avaliação mais detalhada. Durante a leitura dos textos foram separados, de cada artigo, os dados considerados mais relevantes de acordo com os propósitos desta pesquisa. Uma visão geral desses dados é apresentada na próxima seção.

Quadro 1: Total de artigos por tema

TEMA	QTDE
Ensino de IA na Educação Básica	1
Ensino de IA por meio da robótica na graduação em computação	11
Ensino de IA por meio da robótica na graduação - Abordagem baseada em agentes	4
Ensino de IA pra leigos por meio da robótica	3

3. IA no Ensino Fundamental: Temas Norteadores

3.1. IA na Educação Básica

A busca realizada nesta revisão retornou apenas um artigo referente ao ensino de IA na Educação Básica: *Artificial intelligence and computer science in education: From kindergarten to university* [Kandlhofer *et al*. 2016]. O artigo apresenta uma proposta de alfabetização em IA que se estende desde a Educação Infantil até a Universidade e baseia sua metodologia em uma analogia com a alfabetização em leitura e escrita, desenvolvendo módulos distintos para faixas etárias distintas, seguindo a estrutura de "*multi-stage view on reading/writing*" proposta por Neuman *et al.* (2000).

Em analogia às fases de progressão no trabalho de alfabetização apontadas no trabalho de Neuman *et al.* (2000), Kandlhofer *et al.* (2016) propõem o ensino de IA, a partir da Educação Infantil, seguindo os seguintes passos: 1) Conscientização e exploração lúdica de tópicos em IA (Educação Infantil e Ensino Fundamental I). 2) Experimentação e familiarização com as teorias por trás de determinados tópicos de IA; desenvolvendo trabalhos de forma independente e solucionando problemas (Ensino Fundamental II). 3) Promoção de temas centrais em IA e familiarização com tópicos avançados; aquisição e aplicação independente de conhecimento (Ensino Médio). 4) Tornando-se fluente em IA: aplicação de métodos de solução de problemas em níveis mais altos de abstração; promovendo a "compreensão fundamental" de tópicos em IA (Universidade) [Kandlhofer *et al*. 2016].

A partir destas premissas, foram desenvolvidos módulos específicos para cada seguimento escolar. Estes módulos abrangem os seguintes temas, baseados no trabalho de Russell e Norvig (2003): autômatos, agentes inteligentes, grafos e estruturas de dados, ordenação, solução de problemas por busca, planejamento clássico (modelagem de

problemas, tomada de decisão, estabelecimento e avaliação de planos) e lógica, aprendizado de máquina [Kandlhofer *et al*. 2016]. Estes temas são revisitados, com diferentes aprofundamentos e distintas abordagens, em cada seguimento de aprendizagem.

No que se refere à Educação Infantil e Ensino Fundamental I, lança-se mão de técnicas como *storytelling*, Computação Desplugada [Bell *et al*. 2009] e Robótica Educacional para desenvolver as seguintes atividades: grafos e estruturas de dados, algoritmos de ordenação e solução de problemas com busca. Já, para o Ensino Fundamental II, faz-se uso do kit NXT da Lego Mindstorms[3], papel, caneta e Computação Desplugada para se trabalhar, entre outras, as seguintes atividades: maior conscientização acerca dos tópicos de estudo, apresentação de grafos, árvores e estruturas de dados, familiarização com agentes inteligentes por meio da robótica e avaliação de diferentes estratégias/algoritmos de busca. No que tange o uso de robótica para o aprendizado de IA no Ensino Médio são abordados mais especificamente os seguintes conteúdos: autômatos e agentes inteligentes [Kandlhofer *et al.* 2016].

3.2. Ensino de IA via Robótica Educacional na Graduação em Computação

Utilizar a robótica o mais cedo possível no currículo de cursos de IA faz com que os alunos atinjam um alto grau inicial de motivação com o aprendizado, o que acaba por dar sustentação ao interesse desses alunos nos módulos subsequentes [Beer, Chilel e Drushel 1999; Chiou 2002 *apud* Chiou 2012]. O grande desafio nesse sentido é "descobrir" um conjunto de ferramentas (*software + hardware*) que tornem o aprendizado de IA por meio da robótica mais atraente sem que sejam impostas limitações a esse aprendizado. Dentro desse contexto, a facilidade com que os alunos conseguem interagir com os robôs, ou seja, as interfaces de hardware e software adotadas para este fim, assumem fundamental importância [Dodds *et al*. 2006].

Um aspecto que precisa ser levado em consideração em relação ao uso da robótica educacional para o aprendizado de IA é a capacidade de memória e processamento das plataformas de robótica de baixo-custo. Projetos com robótica de baixo custo tendem a se embasar em *local sensing*, sendo o uso de arquiteturas reativas a opção natural neste caso. De acordo com Dodds *et al.* (2006) estas plataformas, muitas vezes, não possuem capacidade computacional suficiente para que se possa implementar, de maneira totalmente embarcada, determinadas tarefas de IA. Greenwald *et al.* (2004), também chamam a atenção para essa questão, sugerindo que implementar algoritmos avançados de IA nessas plataformas pode ser bastante problemático. Essa dificuldade, no entanto, pode ser contornada fazendo-se parte do processo *off-board*, como por exemplo, o aprendizado dos pesos de uma rede neural *backpropagation* [Dodds *et al*. 2006].

Greenwald *et al.* (2006) chamam a atenção para o fato de que o aprendizado de IA com robótica propicia aos alunos a possibilidade de trabalhar no desenvolvimento de soluções completas em IA, uma vez que eles trabalham no planejamento e desenho do sistema, escolha e implementação dos algoritmos e, por fim, análise dos comportamentos do robô por meio da experimentação. Os autores utilizam kits Lego Mindstorms (considerado por eles e por diversos outros autores, como Parsons e Sklar (2004), como sendo kits baratos o suficiente para serem adquiridos mesmo que se esteja trabalhando

[3] https://www.lego.com/en-us/mindstorms

com um orçamento limitado [4]) para trabalhar temas como: busca heurística, aprendizado de máquina, redes bayesianas, redes neurais, filtro de partículas, entre outros. Essas técnicas são utilizadas, por exemplo, para desenvolver atividades nas quais os robôs precisem detectar/desviar de obstáculos ou seguir um foco de luz.

Os kits Lego Mindstorms também são utilizados, há mais de 12 anos, por Irgen-Gioro (2016), para trabalhar, na graduação, tópicos como algoritmos genéticos, redes neurais artificiais e aprendizado por reforço, por meio do desenvolvimento, por exemplo, de carrinhos robô seguidores de linha ou que saibam sair de labirintos. Exercícios simples, mas suficientes para fazer os estudantes entenderem do que se trata IA e, também, que ela está atualmente presente em todos os lugares [Irgen-Gioro 2016].

3.3. Abordagem Baseada em Agentes

Abordar o aprendizado de IA da perspectiva de um agente em seu ambiente é uma metodologia efetiva e simples. Embora os agentes sejam compostos por sensores e atuadores, muitas vezes, em aulas de IA, detalhes acerca desses elementos são menosprezados ou até mesmo ignorados. Já no caso da robótica esses são os conceitos centrais [Parsons e Sklar 2004].

Blank *et al*. (2006) apresentam em seu trabalho o Pyro Toolkit: um conjunto de materiais para o ensino de programação de robôs, tais como, tutoriais, exemplos de paradigmas de programação de robôs, e uma gama de módulos de IA, como por exemplo, computação evolucionária e redes neurais. A ideia do Pyro Toolkit é "promover uma transição suave para o estudante dos agentes simbólicos para os robôs do mundo real, o que reduz significativamente o custo de aprender a usar robôs" [Blank *et al*. 2006, p.39, tradução nossa]. Para tanto, o toolkit introduz "abstrações genéricas de robôs" que servem para possibilitar a programação de uma diversidade de plataformas de robótica por meio da linguagem Python. Desse modo, pode-se programar "comportamentos básicos para os robôs independentemente do tipo, tamanho, peso e formato dos robôs" [Blank *et al*. 2006, p.41, tradução nossa].

O Pyro Toolkit se apropria da abordagem baseada em comportamento para prover ao usuário "unidades" compostas por uma série de comandos que definam um determinado comportamento para os robôs independentemente da plataforma robótica utilizada. A programação do robô é concebida como sendo um conjunto de comportamentos que podem ser disparados a partir de determinadas condições do ambiente [Blank *et al*. 2006]

Parsons e Sklar (2004) apresentam também uma proposta de ensino de IA com base na abordagem baseada em agentes fazendo uso dos kits Lego Mindstorms por meio de atividades tais como: fazer o robô seguir uma linha preta marcada sob uma base branca com aclives e declives, detectar obstáculos usando sensores de toque, reconhecer áreas coloridas, entre outras. São trabalhados, durante o curso proposto, tópicos tais como: introdução à IA, o que são agentes, controle reativo, introdução à robótica, perceptrons, visão de máquina, busca heurística, busca adversarial, representação de conhecimento, lógica proposicional, lógica dos predicados e aprendizado por reforço**.**

[4] Essa avaliação não se adequa a realidade, em especial, das escolas públicas brasileiras. No Brasil, um kit Lego Mindstorms EV3, por exemplo, custa aproximadamente R$ 2.600,00.

3.4. IA para Leigos

Uma das estratégias adotaras neste estudo para a investigação dos possíveis recursos a serem adotados para o aprendizado de IA no Ensino Fundamental, foi a procura por artigos que tratassem do ensino de IA para leigos, entendendo-se por leigos pessoas que não sejam da área de informática, tendo sido selecionados dois artigos sobre esse tema.

No artigo de Gillian e Paradiso (2014), é apresentada uma biblioteca multiplataforma em C++ desenvolvida para tornar o aprendizado de máquina e o reconhecimento de gestos mais acessível a pessoas fora da área de IA. O Gesture Recognition Toolkit (GRT) possui uma série de algoritmos de regressão em tempo real e uma gama de algoritmos para o processamento de sinais e extração de *features*. As principais características do GRT são: acessibilidade (fácil de usar), flexibilidade (multiplataforma e "multipropósitos"), escolha (diversidade de opções de algoritmos), infraestrutura de apoio (diversidade de pré e pós processamentos), customizável (incorporação de novos algoritmos) e suporte a tempo real [Gillian e Paradiso 2014].

O segundo artigo analisado, dentro da temática de IA para leigos, foi o trabalho de Marshall (2004) que apresenta uma proposta de ensino de IA com robótica para alunos de Ciência Cognitiva. O principal objetivo da proposta, no que tange o uso da robótica, é possibilitar aos alunos a percepção de que o comportamento inteligente é intimamente dependente da interação de um sistema físico com o ambiente [Pfeifer e Scheier 1999 *apud* Marshall 2004]. Para tal, são feitos uma série de exercícios baseados nos Veículos de Braitenberg[5] [Braitenberg 1984], que possibilitam, além da percepção já mencionada, o entendimento de que comportamentos aparentemente complexos ou com propósitos "profundos" não exigem, necessariamente, "processos interiores" complexos, ideia esta que tem uma implicação muito forte para o conceito de inteligência [Marshall 2004].

4. Conclusões

Este trabalho tinha por objetivo realizar uma investigação inicial acerca do aprendizado de Inteligência Artificial (IA) no Ensino Fundamental, por meio da Robótica Educacional, bem como de possíveis estratégias a serem adotadas para a implementação desse aprendizado, por considerar-se que esta é uma temática de significativa importância no contexto atual, onde a IA está cada vez mais presente em diversos setores, de maneira que conhecimentos nessa área passam a assumir significativa importância, não só para profissionais das mais diversas áreas como também para toda comunidade.

No que se refere aos trabalhos dedicados especificamente ao aprendizado de IA no Ensino Fundamental, esta investigação apontou indícios da existência de poucas pesquisas acerca desse tema, o que abre um espaço significativo para o desenvolvimento de estudos nessa área. Já no que tange a identificação de possíveis abordagens a serem adotadas para o aprendizado de IA por meio da Robótica Educacional no Ensino Fundamental, partindo-se da avaliação de trabalhos desenvolvidos para o público adulto, identificou-se, em especial, o uso das seguintes estratégias:

---

[5] Um veículo de Braitenberg é um conceito concebido por Valentino Braitemberg. Os movimentos desses veículos são determinados a partir da leitura direta de alguns sensores, sendo que o "comportamento" resultante aparenta ser inteligente, embora seja, de fato, puramente reativo.

- Adoção das abordagens baseadas em agentes e em comportamento em conjunto com arquiteturas puramente reativas. De modo geral utilizando-se as arquiteturas puramente reativas em um primeiro momento e evoluindo-se para soluções mais robustas baseadas no uso de técnicas de IA.
- Uso de carrinhos robóticos equipados com sensores para realização das atividades de aprendizado de IA. Essas atividades geralmente consistem em: seguir uma linha, detectar e desviar de obstáculos, sair de um labirinto, seguir uma fonte de luz e reconhecer cores.
- Utilização de ferramentas que possibilitem aos alunos implementarem soluções baseadas em IA por meio da identificação da técnica mais adequada a ser adotada para a solução de um problema proposto e do entendimento em "alto nível" da lógica de funcionamento dos algoritmos e técnicas utilizados, ou seja, sem a necessidade de conhecer detalhes de "baixo nível" referentes à implementação das técnicas trabalhadas.
- Adoção dos kits Lego Mindstorms para as atividades com robótica.

Observa-se, com base nesses dados, a oportunidade para realização de estudos que busquem desenvolver ferramentas que venham a possibilitar a implantação dessas abordagens com crianças do Ensino Fundamental, bem como permitam determinar se essas abordagens são de fato adequadas ao aprendizado de IA por crianças.

Por fim, pôde-se verificar também, neste estudo, a necessidade de pesquisas que visem identificar técnicas de IA que sejam viáveis de serem implementadas em plataformas de robótica de baixo custo compatíveis com a realidade financeira das escolas da rede pública brasileira, como a plataforma Arduino[6] (em conjunto com sensores e atuadores de baixo custo e embarcadas, por exemplo, em robôs construídos com materiais recicláveis), uma vez que estas plataformas possuem capacidade de memória e processamento bastante reduzidas: uma placa Arduino UNO genérica, no valor aproximado de R$ 45,00, possui 8K de RAM e um processador de 16MHz, enquanto um kit Lego Mindstorms possui, no caso do modelo EV3, 64M de RAM, processador de 300MHz e um preço aproximado de R$ 2.600,00.

5. Trabalhos Futuros

Alguns possíveis trabalhos futuros decorrentes deste estudo são: a realização de uma revisão sistemática sobre os temas aqui explorados e o aprofundamento do estudo de questões identificadas nessa investigação inicial no que concerne a possíveis abordagens a serem adotadas para o aprendizado de IA no Ensino Fundamental. Algumas sugestões de estudos nesse sentido encontram-se na seção 4 deste trabalho.

Referências

Beer, R. D.; Chilel, H. J.; Drushel, R. F. (1999) Using autonomous robotics to teach science and engineering. "Communications of the Association of Computing Machinery", v. 42, n. 6, p. 85-92.

Bell, T. et al. (2009) Computer science unplugged: School students doing real computing without computers. "The New Zealand Journal of Applied Computing and Information Technology", v. 13, n. 1, p. 20-29.

[6] https://www.arduino.cc/

Blank, D. et al. (2006) The Pyro toolkit for AI and robotics. “AI magazine”, v. 27, n. 1, p. 39-50.

Braitenberg, V. (1984) “Vehicles**:** Experiments in Synthetic Psychology.” MIT Press. Cambridge.

Chiou, A. (2002) “Educational robotics: Instructional technology to unify diversity of computing topics into a single cohesive unit”. Proceedings of the Scholarly Inquiry in Flexible Science Teaching and Learning Symposium. Sydney. p. 75-76.

Chiou, A. (2012) “Teaching Technology Using Educational Robotics”. Proceedings of the Australian conference on science and mathematics education (formerly UniServe Science Conference). p. 9-14.

Dodds, Z. et al. (2006) Components, curriculum, and community: Robots and robotics in undergraduate ai education. “AI magazine”, v. 27, n. 1, p. 11-22.

Gillian, N. E.; Paradiso, J. A. (2014) The gesture recognition toolkit. “Journal of Machine Learning Research” , v. 15, n. 1, p. 3483-3487.

Greenwald, L. et al. (2006) Using educational robotics to motivate complete AI solutions. “AI Magazine”, v. 27, n. 1, p. 83-95.

Greenwald, L.; Artz, D. (2004) Teaching Artificial Intelligence with Low-Cost Robots. “Accessible Hands-on Artificial Intelligence and Robotics Education”, p. 35–41.

Irgen-Gioro, J. J. Z. (2016)“Teaching Artificial Intelligence Using Lego”. Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS) p. 209-216.

Kandlhofer, M. et al. (2016) “Artificial intelligence and computer science in education: From kindergarten to university”. Frontiers in Education Conference (FIE). Erie: IEE. p. 1-9.

Marshall, J. B. (2004) “An Introductory CS Course for Cognitive Science Students”. Papers from the 2004 AAAI Spring Symposium.

Neuman, S. B.; Copple, C.; Bredekamp, S. (2000)“Learning to Read and Write: Developmentally Appropriate Practices for Young Children”. Mcgraw-Hill.

Parsons, S.; SKLAR, E. (2004) “Teaching AI using LEGO mindstorms”. AAAI Spring Symposium.

Pfeifer, R.; Scheier, C. (1999) “Understanding Intelligence”. MIT Press. Cambridge.

Queiroz, R. L. (2017) “DuinoBlocks4Kids: utilizando Tecnologia Livre e materiais de baixo custo para o exercício do Pensamentos Computacional no Ensino Fundamental I por meio do aprendizado de programação aliado à Robótica Educacioanal”. Dissertação de Mestrado. UFRJ. Rio de Janeiro. 138 f.

Russell, S. J.; Norvig, P. (2003)“Artificial Intelligence”**:** a Modern Approach. 2 ed. Prentice Hall.

Sklar, E.; Parsons, S. (2002) “RoboCupJunior: A Vehicle for Enhancing Technical Literacy”. AAI Mobile Robot Competition 2002 - Papers from the AAAI Workshop. Edmonton. p. 108 -118.

Geração de Casos de Teste Independentes de Plataforma Utilizando Diagramas de Classes da UML Anotados com Restrições OCL

Marcos V. F. A. Dias[1], Eber A. Schmitz[1], Mônica F. da Silva[1], Priscila M. V. Lima[1]

[1]Universidade Federal do Rio de Janeiro (UFRJ) - Rio de Janeiro, RJ - Brasil

{mvfad@ufrj.br, eber@nce.ufrj.br, monica@nce.ufrj.br, priscila.lima@nce.ufrj.br}

Abstract. *This paper describes a systematic approach for the generation of abstract test cases using the USE tool. The approach comprises a method based on the small scope hypothesis using the USE snapshot generator and the ASSL language. This allows for the generation of test cases through the verification and validation of instances of the conceptual model, where business rules are expressed in the form of UML class diagrams enriched with OCL constraints.*

Resumo. *Este artigo descreve uma abordagem sistemática para geração de casos de teste independentes de plataforma utilizando a ferramenta USE. É apresentado um método baseado na hipótese do escopo reduzido utilizando o gerador de snapshots do USE através da linguagem ASSL, de modo a possibilitar a geração dos casos de teste através da verificação e validação das instâncias do modelo conceitual, com base em regras de negócio expressas em diagramas de classes UML enriquecidos com restrições OCL.*

1. Introdução

O teste de *software* é uma importante parte das atividades de V&V (Verificação e Validação) que visam garantir a qualidade do produto de *software* sendo desenvolvido. Essas atividades avaliam se o sistema em desenvolvimento atende às especificações dos requisitos (funcionais e não-funcionais) como também a satisfação do cliente em relação ao produto [Pressman 2010; Sommerville 2011].

Os maiores esforços gastos durante o ciclo de desenvolvimento de um *software* se concentram na fase de teste, que chegam a atingir cerca de 50% ou mais dos custos de desenvolvimento, representando um grande desperdício de tempo, custo e recursos necessários para produção de um sistema de informação [Warmer e Kleppe 2003; Utting e Legeard 2007; Xu *et al*. 2015].

A dificuldade de garantir a qualidade do *software* cresce consideravelmente à medida que a complexidade dos sistemas aumenta [Silva-de-Souza 2012; Sousa 2009]. As abordagens de desenvolvimento orientadas a modelos (*Model-Driven Software Engineering* - MDSE) permitem concentrar-se na essência do sistema sem que haja qualquer tipo de interferência de tecnologia [Brambilla, Cabot e Wimmer 2012; Jackson, 2006]. O mesmo tipo de abordagem pode ser aplicado à criação de casos de

teste. Na abordagem orientada a modelos uma parte desta complexidade é abstraída do processo de criação de casos de teste.

Em particular, a criação de casos de teste a partir de modelos conceituais que compreendem as entidades do domínio, seus relacionamentos e suas restrições bem definidas, diminui o tempo e os recursos gastos durante o desenvolvimento [Silva-de-Souza 2012]. Além disso, é uma garantia de que o produto vai atender verdadeiramente aos requisitos especificados, sem passar pela interferência da interpretação do analista de teste [Utting e Legeard 2007].

Este trabalho apresenta uma abordagem para a geração de casos de teste abstratos a partir de modelos de classes UML anotados com restrições OCL. A utilização da OCL possibilita especificar regras que não podem ser definidas graficamente no modelo. A abordagem aplica a estratégia da hipótese do escopo reduzido (S*mall Scope Hypothesis*) proposta por [Jackson 2006] ao método *lightweight* de validação de modelos conceituais proposto por [Richters e Gogolla 2005]. A abordagem utiliza a ferramenta USE e sua linguagem para validação de modelos conceituais ASSL.

2. Arcabouço Conceitual

2.1. Teste de *Software*

Teste de *software* é uma atividade do processo de desenvolvimento de *software* que consiste na execução controlada de programas com o objetivo de verificar se o sistema atende aos objetivos a que foi proposto. O teste de *software* visa garantir que o produto sendo desenvolvido esteja dentro dos padrões da qualidade desejada [Dias-Neto 2008; Pressman 2010; Sommerville 2011]. As atividades de teste podem ser executadas de maneira concomitante às atividades do ciclo de vida do desenvolvimento do software [Bastos *et al* 2012].

Um caso de teste descreve uma situação de teste contendo dados de entrada, condições de execução e resultado da saída esperada provida por um oráculo [IEEE 2008]. Uma definição mais simples é vista em [Naik e Tripathy 2008], que apresenta caso de teste apenas como uma dupla contendo dados de entrada e saída esperada.

Um caso de teste pode ser classificado como positivo ou negativo. Um caso de teste positivo é aquele que é projetado de forma a verificar se a resposta do sistema está de acordo com aquela especificada nos requisitos. Um caso de teste negativo, por outro lado, é projetado para verificar a resposta do sistema a situações que estão fora da especificação e, portanto, devem retornar uma situação de erro [Silva-de-Souza 2012; Spillner, Linz e Schaefer 2014]. Casos de teste negativos também estão ligados ao teste de robustez do sistema, uma vez que é necessário que o sistema seja capaz de lidar com situações não previstas (dados de entrada não esperados) sem que haja interrupção completa do mesmo [Spillner, Linz e Schaefer 2014].

A tarefa de especificar (prever) o resultado esperado da execução de um caso de teste é denominada “problema do oráculo”. Em abordagens manuais esta atividade é executada pelo analista de teste que visualiza o resultado da execução de um caso de teste e observa se um teste apresentou sucesso ou erro [Peteres e Parnas 1994]. Dessa maneira, um oráculo é um mecanismo pelo qual se obtém a saída esperada para a

execução de um programa de teste, podendo ser a visualização do analista de teste, um processo, um programa ou mesmo um conjunto de dados [Naik e Tripathy 2008].

2.2. Validação de modelos de domínio

As técnicas de validação *lightweight* representam uma maneira mais simples de se avaliar a consistência das especificações de um domínio se comparada às estratégias tradicionais de validação baseadas em provas matemáticas (*heavyweight*) [Jackson, 2006]. Neste trabalho é abordado algumas formas de validação que se relacionam com as técnicas *lightweight:* a hipótese do escopo reduzido, a ferramenta *UML based Specification Environment* (USE) e a linguagem *A Snapshot Sequence Language* (ASSL).

A hipótese do escopo reduzido sugere que a maioria absoluta das situações de falha que comprometem a execução de um software pode ser replicada com um número pequeno de instâncias dos objetos participantes (*small scope*). Como consequência, se todos os estados possíveis, cada um deles com um número reduzido de objetos, forem avaliados, a grande maioria das falhas será encontrada. Cada falha encontrada mostra um contra-exemplo para as restrições do sistema contidas na sua especificação [Jackson 2006].

A USE é uma ferramenta utilizada para especificação, análise e validação do modelo conceitual do domínio. Ela tem como base a descrição textual do modelo de classes contendo classes, atributos, operações, relacionamentos e restrições descritas em *Object Constraint Language* (OCL) sobre os objetos do domínio. A USE trata o modelo UML de maneira executável, fornecendo animações através de *snapshots*, que são instâncias do modelo conceitual que representam um determinado estado do sistema em um instante do tempo. Essas animações têm por objetivo facilitar a validação do modelo pelo desenvolvedor (*lightweight formal method*) [Richters e Gogolla 2005].

A OCL [OMG 2014] é uma linguagem textual de restrição de objetos definida pela *Object Management Group* (OMG) cujo objetivo é especificar regras que não podem ser definidas graficamente em um modelo aderente ao padrão *Meta Object Facility* (MOF). Desse modo, a OCL permite que as regras de negócio sejam expressas em uma linguagem computacional como um complemento sobre dos aspectos de diagramação da UML. Os principais propósitos da linguagem são especificar restrições através de: invariantes, pré/pós condições sobre as operações, condições de guarda e regras de derivação.

A ASSL é uma linguagem que permite a validação e manipulação de objetos de modelos conceituais especificados através da ferramenta USE. Um *script* ASSL contém uma sequência de instruções responsável por gerar *snapshots* [Richters e Gogolla 2005]. Um *snapshot* é visualizado no USE através de um diagrama de objetos. A figura 1 apresenta um exemplo de *script* ASSL.

```
procedure generatePersons(count:Integer)
var thePersons:Sequence(Person);
begin
thePersons:=CreateN(Person,[count]);
for p:Person in [thePersons]
  begin
  [p].name:=
    Any([Sequence{'Ada', 'Bob', 'Cher', 'Dan', 'Eva', 'Fred'}
    ->reject(n1|Person.allInstances.name->exists(n2|n1=n2))]);
  end;
end;
```

Figura 1: Exemplo de *script* ASSL

A instrução *Try* do ASSL permite que todas as possibilidades de relacionamentos entre os objetos sejam geradas pela ferramenta USE durante o processo de validação da especificação. Contudo, esse processo de validação é interrompido quando um estado inválido é encontrado. Diante disso, foi necessário alterar o código interno da USE para que essa interrupção não ocorra. Desse modo é possível obter os casos de teste negativos.

3. Abordagem Proposta

A abordagem proposta neste trabalho é baseada nos conceitos empregados em Testes Baseados em Modelos, onde o processo de teste de *software* é realizado por intermédio da construção de modelos de teste que representam o estado estrutural e comportamental de um Sistema Sob Teste (SST). Neste caso, foi definido o modelo de teste como sendo a própria especificação do sistema. A figura 2 apresenta as atividades necessárias para realização do processo da abordagem.

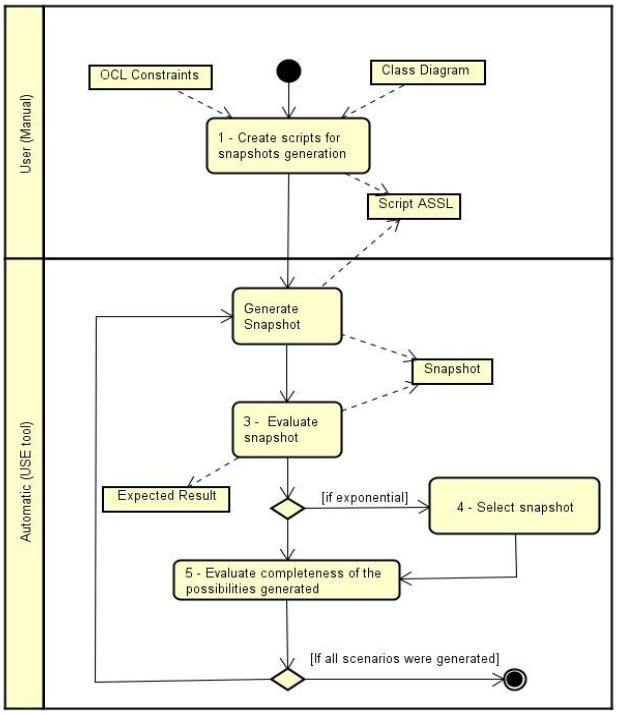


Figura 2: Processo de execução da abordagem proposta

No contexto deste trabalho foi definido um caso de teste como sendo uma dupla (Estado, Situação), onde Estado *(snapshot)* é um conjunto de objetos e associações do modelo conceitual e Situação representa a avaliação desse estado. Um estado é avaliado como válido quando não viola nenhuma das restrições de integridade do modelo e inválido no caso contrário. A figura 3 apresenta os artefatos comuns à abordagem.

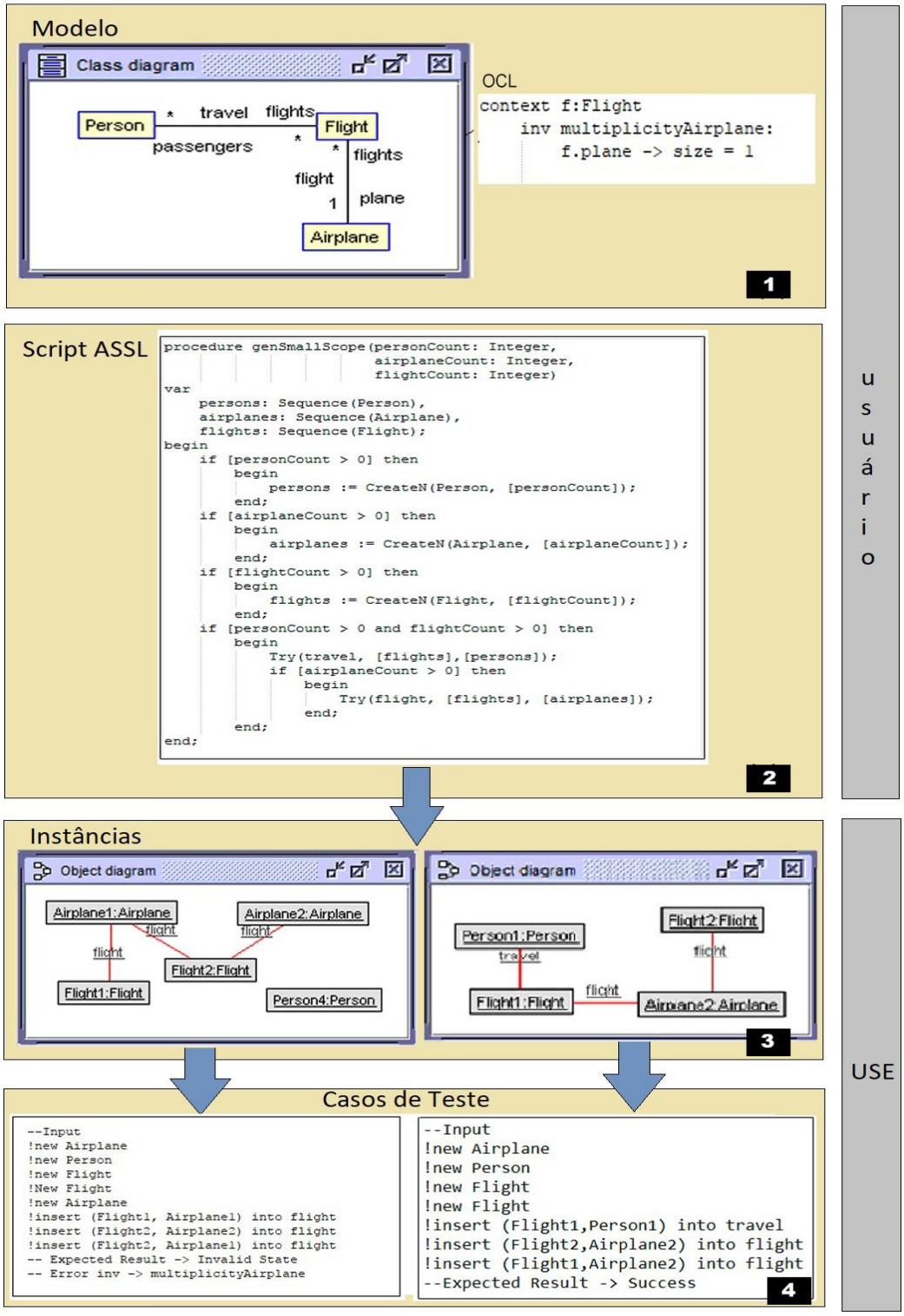


Figura 3: Visão geral dos artefatos que compõem a abordagem

Em (1) é demonstrado o modelo de domínio e sua respectiva restrição de multiplicidade em formato OCL. Em seguida, (2) um *script* ASSL é criado manualmente pelo usuário. Esse *script* deve conter o procedimento necessário para execução da hipótese do escopo reduzido. O passo (3) apresenta duas instâncias do modelo que foram geradas durante a execução do *script* ASSL. E por fim, (4) demonstra dois possíveis casos de teste gerados, um negativo e outro positivo.

Quando a geração dos casos de teste se tornar inviável devido ao crescimento exponencial do número de *snapshots* gerados, é empregado uma técnica de seleção baseada na combinação das restrições violadas, de modo que os casos sejam armazenados com até dois objetos (3 possibilidades – zero, um ou dois obejtos) de cada violação. Então, um modelo contendo *n* invariantes OCL irá gerar *snapshots* até que os 3^n casos de teste negativos sejam encontrados.

4. Comparativo com Trabalhos Relacionados

Diversas abordagens para geração de casos de teste com base em instâncias do modelo conceitual UML já foram descritas na literatura [Bizerra *et al*. 2012]. Contudo, grande parte dessas propostas não apresenta ligações com as regras de negócio descritas em linguagem OCL e não utilizam a ferramenta USE como parte fundamental para geração de casos de teste para o SST. Nesse aspecto, os trabalhos que apresentam maiores semelhanças a este são os de Araújo (2009) e Bizerra *et al* (2012).

Araújo (2009) propõe um método para validar a conformidade dos processos de negócio em relação às regras de negócio, onde a validação do modelo é baseada na animação de um conjunto de cenários executados através da ferramenta USE. No trabalho de Araújo, o analista também é responsável pela especificação dos *scripts* ASSL que serão utilizados para geração dos cenários de teste. No entanto, a estratégia utilizada por ele é baseada em caminhos chaves de um diagrama de atividades, enquanto que este trabalho emprega estratégia baseada no número de objetos participantes do cenário, sendo necessária apenas a utilização do diagrama de classes.

O trabalho de Bizerra *et al*. (2012) propõe um método para geração de instâncias de teste com base apenas em diagrama de classes e OCL, porém esse trabalho não especifica qual é a estratégia utilizada para a geração dos cenários com relação às validações de integridade referencial do modelo, limitando-se a descrever as estratégias para geração de valores para atributos e pré e pós-condições sobre as operações do modelo. A proposta de Bizerra *et al*. (2012) é apenas para validar modelos, onde o oráculo para os casos de teste é o analista que observa a animação e a valida, ao contrário deste trabalho onde o oráculo de teste é automatizado com base na execução através da ferramenta USE.

O diferencial deste trabalho em relação aos demais consiste na utilização de um único diagrama da UML para geração dos casos de testes e, principalmente, da geração automática do oráculo de teste a partir da utilização de métodos de validação *lightweight* da especificação do sistema. Além disso, os casos de teste gerados pela abordagem podem ser posteriormente transformados em testes específicos de alguma plataforma.

5. Considerações finais

Este trabalho apresentou uma abordagem sistematizada para a geração de casos de teste independentes de plataforma. A proposta apresentada segue os princípios dos métodos *lightweight*, usados na verificação de modelos, para a geração dos casos de teste funcionais sobre as restrições do modelo.

Por intermédio da utilização de diagramas de classes anotados com restrições de negócio em OCL é possível gerar casos de teste juntamente com os oráculos de teste a partir da validação da especificação do *software*. Os casos de teste gerados a partir de uma especificação correta do sistema garantem que os requisitos funcionais serão testados sem os problemas causados pela má interpretação dos requisitos por parte de um analista de teste trabalhando de forma independente.

O analista de teste, muitas vezes, não está tão familiarizado com o domínio do sistema como o analista especializado em modelagem de requisitos do *software*. O problema dos erros de interpretação é minimizado quando as especificações dos testes passam a ser de responsabilidade do analista que especificou o sistema. A abordagem proposta neste trabalho possibilita que os analistas de requisitos que saibam modelar corretamente sistemas utilizando UML e OCL, além da ASSL realizem a especificação dos testes funcionais para um sistema.

5.1 Trabalhos Futuros

A atividade de criação manual dos *scripts* ASSL demanda conhecimento a respeito da linguagem e ainda é propensa a alguns erros. Contudo, essa etapa poderia ser automatizada em trabalhos futuros. Além disso, a linguagem ASSL é baseada em um algoritmo enumerativo, onde cada *snapshot* é feito sequencialmente pela USE em um processo que pode ser considerado lento. A geração de *snapshots* para modelos contendo muitas classes e relacionamentos pode se tornar inviável devido ao crescimento exponencial ocasionado pela hipótese do escopo reduzido. Trabalhos futuros podem tratar da seleção de um conjunto ótimo de casos de teste, a fim de que o processo se torne mais veloz.

Agradecimentos

Nossos agradecimentos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) pelo auxilio através de bolsa de estudos para prosseguimento deste trabalho.

Referências

Araujo, M. B. Um método para validar a conformidade de processos de negócio com regras de negócio. Dissertação (Mestrado em Informática) – Programa de Pós-Graduação em Informática – Núcleo de Computação Eletrônica – UFRJ, Rio de Janeiro, Brasil, 2010.

Bastos, A.; Rios, E.; Cristalli, R. Moreira, T. Base de conhecimento em teste de software. São Paulo, Martins Fontes, 2012.

Bizerra, E., Silveira, D., Cruz, M., Wanderley, F. A method for generation of tests instances of models from business rules expressed in OCL. IEEE Latin America Transactions, 10(5), 2105-2111, 2012.

Brambilla, M., Cabot, J., Wimmer, M.: Model-Driven Software Engineering in Practice. Morgan & Claypool Publishers, 2012.

Dias-Neto, A C. Introdução a Teste de Software. Engenharia de Software Magazine nº1, 2008.

IEEE 829. Standard for Software Test Documentation - Description, ANSI/IEEE 829-2008, 2008.

Jackson, D. Software abstractions: logic, language and analysis. MIT Press, Cambridge, MA, 2006.

Naik, k. Tripathy, P. Software Testing and Quality Assurance: Theory and Practice. Wiley. A John Wiley & Sons, inc., Publication, 2008

Object Management Group (OMG). Object Constraint Language (OCL), Specification Version 2.4. OMG Document No. formal/2014-02-03, http://www.omg.org/spec/OCL/2.4/PDF, Fevereiro 2014.

Peters, D. Parnas, D. Generating a Test Oracle from Program Documentation. Proceeding ISSTA '94 Proceedings of the 1994 ACM SIGSOFT international symposium on Software testing and analysis Pages 58-65. 1994.

Pressman, R., Software Engineering: A Practitioner's Approach. 7th ed. New York: McGraw-Hill, 2010.

Richters, M. Gogolla, M. Validating UML models and OCL constraints, University of Bremen, FB3, Computer Science Department, Germany. 2005.

Silva-de-Souza, T. Uma Abordagem Baseada em Especificação para Teste de Web Services RESTful. 2012 Dissertação (Mestrado em Informática) - Programa de Pós-graduação em Informática – Núcleo de Computação Eletrônica - Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

Spillner, A. Linz, T. Schaefer, H. Software Testing Foundations: A Study Guide for the Certified Tester Exam. Foundation Level ISTQB compliant. 4.ed. 2014.

Sommerville, I. Engenharia de software. 9. ed.São Paulo: Pearson, 2011.

Sousa, H. Construção Automatizada de Casos de Teste Usando Engenharia Dirigida por Modelos. 2009. Dissertação (Mestrado em Engenharia de Eletrecidade) – Programa de Pós-Graduação em Engenharia de Eletrecidade – Universidade Federal do Maranhão. Maranhão, Brasil.

Utting, M. Legeard, B. Pratical Model-Based Testing: A Tools Approach. Elsevier / Morgan Kaufmann. 2007.

Warmer, J., Kleppe, A.: The Object Constraint Language, The: Getting Your Models Ready for MDA. 2th. Addison Wesley, 2003.

Xu, W. Michael Kent, Lijo Thomas, and Linzhang Wang; An Automated Test Generation Technique for Software Quality Assurance IEEE Transactions on Reliability. 2015.

Sistema Computacional de Monitoramento de Qualidade de Água baseado em Arduino

Felipe Schubert Costa[1], Lucas F. Pinheiro[1], Roberto S. G. Pontes[1], Diego Brandão[1], Gabriel R. Gomes[1], Gabriel dos S. L. Stefano[1], Henrique M. A. Junior[1], Raphael O. Guerra[2]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ) – Nova Iguaçu, RJ – Brasil

[2]Instituto de Computação
Universidade Federal Fluminense (UFF) – Niterói, RJ – Brasil

diego.brandao@eic.cefet-rj.br, rguerra@ic.uff.br

Abstract. *The preservation of water is fundamental for the maintenance of life. Society has a fundamental role to inspect and control the quality of water to keep it fit for consumption. The present work consists of the development of a low cost system to monitor the physical-chemical properties of water in order to determine its quality. An infrastructure of wireless sensor networks (WSNs) has been developed for this. In this WSN each node monitors a set of water properties (such as pH, dissolved oxygen and temperature). The data collected by the network is made available on the web allowing users to view and retrieve data for study and analysis.*

Resumo. *A preservação da água é fundamental para a manutenção da vida. Cabe à sociedade o papel de fiscalizar e controlar a qualidade da água para mantê-la própria para consumo. O presente trabalho consiste no desenvolvimento de um sistema de custo reduzido para monitorar propriedades físico-químicas da água visando determinação da sua qualidade. Uma infraestrutura de redes de sensores sem fio (RSSF) foi desenvolvida para isso. Nessa RSSF cada nó monitora um conjunto de propriedades da água (entre pH, oxigênio dissolvido e temperatura). Os dados coletados pela rede são disponibilizados na web, possibilitando a usuários sua visualização para estudos e análise.*

1. Introdução

A existência e manutenção da vida são impossíveis sem a existência da água. Segundo dados da Agência Nacional de Águas (ANA), o Brasil é o país que possui o mais importante patrimônio hídrico do planeta, cerca de 12% da água doce superficial do mundo, sendo a região Amazônica detentora de 73,6% dos recursos hídricos superficiais nacionais [ANA 2007]. Dessa forma, uma das principais missões do país é a de proteger, preservar e desenvolver condições para manutenção da qualidade e disponibilidade da água.

Nesse contexto, o país tem a Lei n. 9.433/97, a chamada política Nacional de Recursos Hídricos. Dentre os principais fundamentos dessa lei podemos destacar que a água é tida como um recurso finito, de domínio público e valor econômico definido, assim sua gestão deve requerer a participação de toda a sociedade [Kelman 1999].

O monitoramento do corpo hídrico fornece subsídios que permitem o desenvolvimento de técnicas visando determinar o melhor uso da água, sua qualidade, preservação e a compreensão do comportamento de qualquer constituinte lançado nele. Entre os indicadores químicos da qualidade da água, destacam-se: condutividade, taxa de oxigênio dissolvido, demanda bioquímica de oxigênio (DBO), pH (índice de acidez ou alcalinidade), concentração de nitrogênio e fósforo [CNPq 2013].

A utilização de sensores aplicados no monitoramento ambiental cresceu consideravelmente nos últimos anos [Gertz 2012, Urso et. al 2012]. No contexto brasileiro, por exemplo, o estado de São Paulo monitora a qualidade da água desde 2000 [CETESB 2017]. Esse aumento da utilização de sensores no monitoramento ambiental deve-se principalmente pelos avanços tecnológicos e a redução do custo desses equipamentos [CNPq 2013]. Todavia, os valores ainda são altos para a realidade brasileira, o que reduz consideravelmente a sua aplicabilidade. No caso do estado de São Paulo, por exemplo, apesar de possuir mais de 400 pontos de monitoramento, apenas 16 são sistemas automáticos, ou seja, menos de 4% dos pontos [CETESB 2017].

Os nós que constituem a rede de sensores são formados por unidades de processamento associados com módulos de comunicação e os coletores de dados (sensores físico-químicos). Com relação à unidade de processamento, diversos dispositivos podem ser utilizados dependendo da aplicação, do ambiente e do orçamento. Uma opção relativamente econômica, funcional e de simples implementação envolve a utilização da plataforma Arduino, uma solução de *open-hardware* muito difundida na literatura, disponível em diversos modelos e compatível com a grande maioria dos sensores e atuadores disponíveis no mercado.

As redes de sensores sem fio (RSSF) têm diversas aplicações na atualidade sendo o monitoramento ambiental uma delas. Nesta linha existem diversos sistemas propostos, entretanto estes são concebidos para um funcionamento *ad-hoc* ficando assim restrito a um nicho específico. O'Flyrm et al. (2007) desenvolveram um sistema para monitorar propriedades da água de acordo com diretrizes da União Europeia (UE). Eles utilizaram o protocolo ZigBee padrão para comunicação entre os nós sensores, além de utilizar o sistema Tyndall [T.N. Institute 2013]. O sistema desenvolvido verificava as propriedades de temperatura, pH, fosfato, oxigênio dissolvido, condutividade elétrica, turbidez e nível da água. Já o trabalho de Dinh et al. (2007) utilizava uma RSSF para medir a salinidade e profundidade da coluna de água. Eles

utilizavam uma plataforma Fleck3 [CSIRO 2013]. O trabalho de Rao et al. (2013) utiliza o Arduino Mega em um sistema embarcado para realizar o monitoramento do corpo hídrico. O sistema desenvolvido por Rao et al. (2013) difere do apresentado neste artigo, pois, além dos autores citados utilizarem uma versão mais robusta de Arduino no seu trabalho, necessitam de uma interação direta com o sistema para coleta dos dados, uma vez que não dispõe de uma infraestrutura de RSSF para transporte das informações. O trabalho de Faustine et al. (2014) utiliza a mesma plataforma, contudo um módulo Bluetooth permite que os dados monitorados sejam enviados para um aparelho smartphone onde um aplicativo disponibiliza as informações na *web*. Outra variação desse sistema utilizando o protocolo ZigBee para a comunicação é apresentado por Jin-feng et al. (2015). Todos esses trabalhos utilizam a plataforma Arduino Mega para implementar o sistema de sensores. Apesar de facilitar o desenvolvimento do sistema ele aumenta o custo do projeto e por isso no presente trabalho foi utilizado somente o microcontrolador da plataforma básica Arduino UNO R3, conhecido como ATMEGA328.

Este trabalho apresenta um sistema de monitoramento das propriedades da água com um custo menor do que as soluções de mercado, conforme verificado na seção de resultados, permitindo assim um monitoramento mais abrangente.

O presente artigo está estruturado da seguinte maneira: a próxima seção apresenta a arquitetura do sistema desenvolvido. Na seção 3 são apontados alguns resultados demonstrando o funcionamento do sistema. As considerações finais e trabalhos futuros são apresentados na seção 4.

2. O Sistema Pro+

O sistema desenvolvido foi denominado pelos desenvolvedores de ProMAIS (Projeto de Monitoramento Ambiental utilizando Arduino Integrado com Sensores) e posteriormente teve o nome estilizado para Pro+. Este pode ser subdividido em dois conjuntos de elementos: *software* e *hardware*.

Foi desenvolvido um *software* para permitir a visualização de dados através de interface *web*. Este aplicativo foi programado em linguagem Python com o uso do *framework* para *web* denominado Django. Esta opção foi feita pelo rápido desenvolvimento, segurança e escalabilidade no *framework*. A modelagem do banco de dados implementado no Pro+ foi estabelecida utilizando o mapa objeto relacional da ferramenta fazendo com que o modelo fosse construído a partir das classes. Algumas classes, geradas a partir de funcionalidades nativas do framework, tem as tabelas a elas relacionadas prefixadas, por padrão, com a expressão “django”. A gestão de usuários e sessões dos mesmos é feita através de funcionalidade do *framework* denominada *authentication* sendo a implementação totalmente definida pelos *patterns* adotados pela equipe de desenvolvimento do Django.

A interface do sistema *web* do sistema permite a visualização dos dados (identificação e valores medidos) de cada nó individualmente. Para ter acesso a interface principal do sistema é necessário que o usuário realize um cadastro. Usuários com mais privilégios podem criar novos nós, bem como excluir os já existentes.

O sistema permite aos seus usuários que diversas características da rede de sensores sejam visualizadas, não somente sobre as propriedades que estão sendo medidas, mas também a posição geográfica dos nós.

O *hardware* desenvolvido teve seu controlador baseado na plataforma Arduino UNO R3, onde cada nó do sistema consiste em:

- Sensor de pH: Responsável por medir o nível de acidez da água. Escolhido pois dependendo do nível que se encontra pode alterar o metabolismo de espécies aquáticas e também potencializar o efeito de substâncias tóxicas na água. O sensor utilizado capta valores entre 0 e 14, funcionando em seu pleno estado em temperaturas entre +1° C e +99°C.
- Sensor de Temperatura: Medidor de temperatura da água. Selecionado devido à influência em outras propriedades da água, como a condutividade elétrica que é outra característica medida neste trabalho. O instrumento utilizado possui uma faixa de temperatura entre -55 °C e +125 °C, tendo uma precisão de ± 0,5 °C entre -10 °C e +85 °C.
- Sensor de Oxigênio Dissolvido (OD): Possui a função de medir a concentração de oxigênio na água. É de suma importância a sua medição pois grande parte dos seres marinhos precisam de oxigênio para sobreviver, além de ter a função de oxidar material orgânico presente na água. Tal equipamento opera com 2,5V até 5,5V, possuindo uma precisão de 3 casas decimais e fornecer as medidas em miligrama por litro (mg/L).

Em síntese, a escolha destes sensores permite a modelagem matemática que descreve com simplicidade a qualidade da água. Os sensores de pH e de OD são comerciais da Atlas Scientific.

Além dos sensores os nós são compostos por um GPS e um módulo para comunicação sem fio. O GPS utilizado foi o EM-411 SiRF Star III Chipset.

Nos módulos desenvolvidos foi utilizado o modelo de transmissor sem fio Digi Xbee S1 que opera a uma frequência de 2.4GHz sob o protocolo *peer-to-peer* IEEE 802.15.4, por padrão. Para alcançar o funcionamento em rede do tipo *mesh* foi realizado o procedimento de atualização do *firmware* para operar com o protocolo Digimesh, fornecido pela fabricante. A principal vantagem desse protocolo perante outros no mercado é que possibilita a todos os nós da rede adormecerem ciclicamente (de forma síncrona ou assíncrona) enquanto não transmitem dados, economizando assim energia.

Os experimentos aqui apresentados foram inicialmente construídos utilizando uma protoboard e o Arduino UNO R3. Esta estratégia permitiu variar os circuitos de maneira a obter o melhor resultado para desenvolver o módulo final, que foi criado fazendo uso apenas do microcontrolador do Arduino UNO R3, o ATMEGA328. Esta unidade de processamento pertence à família de dispositivos AVR desenvolvida pela empresa Atmel. É um dispositivo RISC de chip único que implementa a arquitetura Harvard de 8-bit (μC).

O nó apresentado neste artigo encontra-se esquematizado pelo diagrama em blocos da Figura 1. Ele possui um algoritmo que gerencia as transmissões dos pacotes de dados captados pelos sensores, além de buscar a melhor rota possível para o tráfego de dados, poupando assim energia. O circuito aguarda a informação dos pacotes de

dados oriundos da porta serial, e estes são então armazenados em um vetor de caracteres. Quando o vetor está pronto, o algoritmo inicia uma rotina para compilar todas as leituras e enviar os dados concatenados para um módulo de comunicação sem fio. Este, por sua vez, faz o envio dos dados para os outros módulos e consequentemente para o *gateway*. O esquema de funcionamento da rede proposta é exemplificado na Figura 2 onde é demonstrada a rede com mais de um nó em operação simulando um cenário de mudança de roteamento por queda de um nó de repasse.

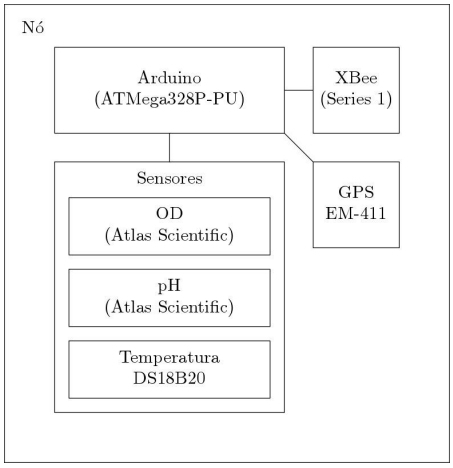


Figura 1. Nó em esquema modular

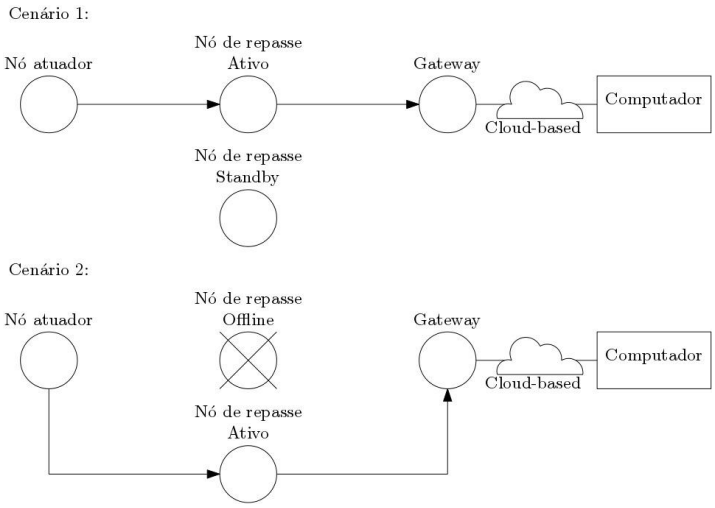


Figura 2. Esquema de funcionamento da infraestrutura de comunicação

O nó foi encapsulado de forma a permitir o contato do material em análise com os sensores, porém impedir que detritos ou materiais com potencial de danos por impacto ao corpo dos sensores consigam alcançá-los. Isso está evidenciado na Figura 3.

Figura 3. Protótipo de um nó do sistema

3. Resultados

A Tabela 1 apresenta um resumo do consumo médio de energia do sistema desenvolvido durante um período de pleno funcionamento. O consumo ainda é alto para ser integrado com células fotovoltaicas, mas está dentro do gerado por pilhas alcalinas, viabilizando sua utilização em aplicações reais.

Módulo	Consumo [mA]
Gateway (em espera)	200
Arduino	60
XBee	32,2
Total do Sistema	292,2

Tabela 1. Consumo de energia do sistema Pro+

A figura 4 apresenta os dados coletados pelo sistema em simulações realizadas em laboratório. Testes em ambientes reais e com maior duração deverão ser realizados. Todavia o gráfico demonstra que o envio dos dados coletados pela rede para o sistema *web* ocorreu de forma perfeita.

No que tange a redução de custo, uma das propostas do projeto, o preço estimado no ano de 2013, desconsiderando valores de frete e de importação, do *hardware* do sistema, para operar com um nó e um *gateway*, foi de cerca de R$ 1600. O valor pode até parecer alto quando visto sem referência, porém se comparado às sondas de mercado observa-se que é consideravelmente vantajoso. Em comparação, realizada através de pesquisa de preço, com a adição de dois sensores, que deixariam o custo da solução em torno de R$ 2200, – considerando o ano supracitado, e também desconsiderando valores de frete e de importação – esta teria características próximas à uma sonda comercial que pode facilmente ultrapassar o valor de R$10000 e somente oferece suporte para leituras locais [TRACOM 2017]. Com isso ficou evidente aos autores o êxito no objetivo de redução de custo, mesmo que estes valores ainda sejam relativamente altos.

Figura 4. Interface do sistema mostrando leituras executadas por um nó protótipo.

4. Considerações Finais e Trabalhos Futuros

Este projeto apresenta uma alternativa de sistema para monitoramento de qualidade de água. O sistema apresenta um custo reduzido quando comparado com as soluções de mercado. Os dados monitorados podem ser visualizados por um aplicativo *web*, e nele são apresentados gráficos exibindo a variação da propriedade selecionada no decorrer do tempo. O sistema atual consiste em dois nós que monitoram as seguintes propriedades da água: pH, condutividade elétrica, temperatura e oxigênio dissolvido. Além disso, as leituras fornecem a posição geográfica do nó obtida via GPS. A característica modular do sistema permite que novos nós sejam integrados à rede, acarretando na mudança da topologia de rede *peer-to-peer* para uma rede *mesh*.

O sistema utiliza o protocolo Digimesh de comunicação sem fio. Todavia, novos protocolos deverão ser avaliados visando uma maior economia de energia. Abordagens baseadas em energia solar deverão ser avaliadas, bem como a transformação da energia gerada pelo fluxo de água em energia elétrica para alimentar os nós-sensores.

Uma nova arquitetura baseada no mapeamento dos nós em sensores virtuais também será avaliada. Espera-se com essa abordagem possibilitar a substituição dos sensores físicos sem comprometer a confiabilidade do sistema, uma vez que a acurácia e precisão das leituras estão intimamente ligadas ao sensor utilizado. Políticas de segurança da informação que trafega na rede também serão avaliadas. Novos filtros para retirada dos ruídos dos sensores serão implementados. Por fim, espera-se que o sistema seja integrado a um simulador hidrodinâmico, fornecendo uma ferramenta que permite a gestores avaliarem derramamentos de poluentes, bem como o perigo de enchentes decorrentes do excesso de chuvas.

Agradecimentos

Os autores agradecem ao programa de extensão do CEFET-RJ (DIREX), ao MEC/PROEXT pelo apoio ao projeto intitulado “Projeto para incentivar alunos dos Ensinos Básico e Superior da Baixada Fluminense nas áreas Tecnológicas” e a FAPERJ pelo apoio por meio do edital ADT1 2017.

Referências

ANA (Agência Nacional de Águas). (2007) "Geo Brasil Recursos Hídricos – Componente da Série de Relatórios sobre o Estado e Perspectivas do Meio Ambiente no Brasil", http://www.ceivap.org.br/estudos/GEO-Brasil-Recursos-Hidricos.pdf, Janeiro.

Kelman, J. (1999) "Evolution of Brazil's Water Resources Management System", Em: Water Resources Management. ABRH, Porto Alegre, p. 19-36.

CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). (2013) "Água: Desafios Da Sociedade", http://estatico.cnpq.br/portal/premios/2013/pjc/imagens/publicacoes/01_cadernoProfessor_xxvii_pjc_web.pdf.

Gertz, E. J. (2012) "Environmental Monitoring with Arduino". O'Reilly, CA.

Urso, L., Astrup, P., Helle, K.B., Raskob, W. e Kaiser, J.C. (2012) "Improving evaluation criteria or monitoring networks of weak radioactive plumes after nuclear emergencies". Em: Environmental Modelling & Software. Volume 38, p. 108–116.

CETESB. (2017) "Relatório de Qualidade das Águas Superficiais do Estado de São Paulo", http://www.cetesb.sp.gov.br/agua/aguassuperficiais/35-publica, Agosto.

O'Flyrm, B., Martinez, R., Cleary, J., Slater, C., Regan, F., Diamond, D. e Murphy, H. (2007) "Smartcoast: A wireless sensor network for water quality monitoring". Em: 32nd IEEE Conference on Local Computer Networks. IEEE, p. 815–816.

T. N. Institute. (2013) "Wireless sensor network (WSN) development". Abril.

Dinh, T. L., Hu, W., Sikka, P., Corke, P., Overs, L. e Brosnan, S. (2007) "Design and deployment of a remote robust sensor network: Experiences from an outdoor water quality monitoring network". Em: 32nd IEEE Conference on Local Computer Networks. IEEE, p. 799– 806.

CSIRO. (2013) "Wireless sensor network devices", http://www.ict.csiro.au/page.php?cid=87, Abril.

Rao, A.S., Gubbi, J. e Palaniswami, M. (2013) "Design of Low-Cost Autonomous Water Quality Monitoring System". Em: Advances in Computing, Communications and Informatics (ICACCI). IEEE, Agosto.

Faustine, A., e Mvuma, A. (2014) "Ubiquitous Mobile Sensing for Water Quality Monitoring and Reporting within Lake Victoria Basin". Wireless Sensor Network. 6, p. 257-264.

Jin-feng, L. e Shun, C. (2015) "A Low-cost Wireless Water Quality Auto-monitoring System". International Journal of Online Engineering. Vol. 11 Issue 3, p. 37-41.

TRACOM. (2017). Analisador de Qualidade de Água Disponível em: http://www.tracom.com.br/novo/?pag=produto-visualiza&id=OA==, Outubro

Organização **Realização**

UFRJ

ISBN: 978-85-7669-421-2 (online)