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

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

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

Sociedade Brasileira de Computação (SBC)

Organizadores
Sergio Manuel Serra da Cruz (UFRRJ)
Raimundo José Macário Costa (UFRRJ)

Promoção
Universidade Federal Rural do Rio de Janeiro (UFRRJ)
Programa de Educação Tutorial PET-SI (PET-SI/UFRRJ)

Realização
Sociedade Brasileira de Computação (SBC)

Editores

Sergio Manuel Serra da Cruz (Universidade Federal Rural do Rio de Janeiro)
Raimundo José Macário Costa (Universidade Federal Rural do Rio de Janeiro)

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

Local – Seropédica/RJ, de 25 e 26 de outubro de 2016

Ano de Publicação – 2016

Edição – 1ª

Editora – Sociedade Brasileira de Computação - SBC

Organizadores – Sergio Manuel Serra da Cruz (Universidade Federal Rural do Rio de Janeiro)
Raimundo José Macário Costa (Universidade Federal Rural do Rio de Janeiro)

ISBN: 978-85-7669-356-7

Agência Brasileira do ISBN
ISBN 978-85-7669-356-7
9 788576 693567

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 2016, a Sociedade Brasileira de Computação (SBC), a Universidade Federal Rural do Rio de Janeiro (UFRRJ) e o Programa de Educação Tutorial (PET-SI/UFRRJ) realizaram, entre 25 e 26 de outubro, no Pavilhão de Aulas Práticas do curso de Sistemas de Informação da UFRRJ a III Escola Regional de Sistemas de Informação do Rio de Janeiro (ERSI-RJ 2016).

Durante a III ERSI-RJ 2016, ocorreram diversas atividades, entre painéis, sessões técnicas, minicursos e reuniões, também houveram importantes discussões relacionadas com o futuro dos Currículos de Referência para os cursos de Sistemas de Informação no Brasil. A escola abrigou a primeira edição do Fórum Regional de Sistemas de Informação do Estado do Rio de Janeiro (FrESI-RJ).

O FrESI-RJ é uma iniciativa da Comissão Especial de Sistemas de Informação e da Comissão de Educação da SBC, seu objetivo foi reunir coordenadores de cursos, professores e alunos dos cursos de Sistemas de Informação para a discussão e elaboração conjunta do Currículo de Referência para os Cursos de Graduação em Sistemas de Informação no Brasil, a ser chancelado pela SBC.

A III 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 terceira 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 minicursos em diversas áreas.

Este ano tivemos mais de 850 visitantes únicos no sítio da escola (http://labbd.ufrrj.br/ersi2016/). Além disso, mais de 70 inscrições e 32 trabalhos do tipo completo foram submetidos (oriundos de diversos estados brasileiros), com 15 trabalhos aceitos para serem apresentados. Para que a III 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 III ERSI-RJ.

Agradecemos também à SBC, às redes CYTED (smartlogistics@ib e BigDSSAgro) e a UFRRJ pela oportunidade e apoio para organizar um evento dessa natureza no campus universitário de Seropédica. E, por último, mas não menos importante, agradecemos a todos componentes do grupo PET-SI/UFRRJ pela extraordinária dedicação ao evento e pela competência impar 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 III ERSI-RJ 2016.

Sérgio Manuel Serra da Cruz

Comitê Organizador do Evento

Coordenação Geral

Sergio Manuel Serra da Cruz (UFRRJ)
Raimundo José Macário Costa (UFRRJ)

Coordenação de Programa

Gizelle Kupac (UFRRJ)
Eduardo Kinder (UFRRJ)

Coordenação de Minicursos

Tiago Cruz (UFRRJ)
Luis Fernando Orleans (UFRRJ)

Coordenação das Palestras e Painéis

Gizelle Kupac (UFRRJ)
Sergio Manuel Serra da Cruz (UFRRJ)

Comitê de Programa

Alexandre Correa (UNIRIO)
Alexandre Sena (UERJ e UNILASALLE)
André Luiz de Castro Leal (UFRRJ)
Carlos Eduardo Melo (UFRRJ)
Claudia Cappelli (UNIRIO)
Claver Pari Soto (UFRRJ)
Cristiano Maciel (UFMT)
Daniela Trevisan (UFF)
Diego Brandão (CEFET-RJ)
Ecivaldo Matos (UFBA)
Eduardo Kinder Almentero (UFRRJ)
Fabiana Mendes (UnB)
Fernanda Baião (UNIRIO)
Flávia Maria Santoro (UNIRIO)
Geiza Hamazaki (UNIRIO)
Geraldo Xexéo (UFRJ)
Gizelle Kupac Vianna (UFRRJ)
Gleison Santos (UNIRIO)

Henrique Sousa (PUC-RJ)
Isabel Cafezeiro (UFF)
Jonice Oliveira (UFRJ)
José Ricardo Cereja (UNIRIO)
Jose Viterbo (UFF)
Laci Mary Manhães (UFRJ)
Leonardo Cruz (UFF)
Luis Alfredo (COPPE/UFRJ)
Luis Fernando Orleans (UFRRJ)
Luiz Maltar Castello Branco (UFRRJ)
Márcio Nunes de Miranda (UFRRJ)
Maria Luiza Machado Campos (UFRJ)
Raimundo José Macário Costa (UFRRJ)
Renata Mendes Araújo (UNIRIO)
Rodrigo Salvador Monteiro (UFF)
Rosangela Lima (UFF)
Sergio Crespo (UFF)
Sérgio Manuel Serra da Cruz (UFRRJ)
Sean W. M. Siqueira (UNIRIO)
Simone Bacellar Leal Ferreira (UNIRIO)
Tiago Cruz de França (UFRRJ)
Victor Almeida (Petrobras)

ARTIGOS TÉCNICOS

Trabalhos Premiados – Melhores Artigos

Sessão Técnica 1

Chair. Profa. Claudia Cappelli (UNIRIO)

Sessão Técnica 2

Chair: Prof. José Viterbo (UFF)

Sessão Técnica 3

Chair. Prof. Luis Orleans (UFRRJ)

Sessão Técnica 4

Chair. Prof. Eduardo Kinder (UFRRJ)

O uso de uma Linguagem Cidadã por diversos perfis organizacionais

Luiz Paulo Carvalho[1], Flávia Maria Santoro[1], Claudia Cappelli[1,2]
[1]BSI– Bacharelado em Sistemas de Informação
[2]PPGI– Programa de Pós-Graduação em Informática
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Av. Pasteur, 296 - Urca - Cep 22290-240 – Rio de Janeiro – RJ – Brasil

{luiz.paulo.silva, flavia.santoro, claudia.cappelli}@uniriotec.br

Abstract. *Business process models are artifacts that represents essential information about an organization and its operation model; however, they are not always understandable by the people interested and involved with them, due to languages and formal technical notations required to capture and conceptually represent their data. A Citizen Language used to give a best understands for citizens can allow a best comprehension also to other roles involved in the process. This paper presents an extension proposal for a Citizen Language in order to improve understanding of other roles involved with organizational processes.*

Resumo. *Modelos de processos de negócios são artefatos que representam informações essenciais sobre uma organização e seu modo de operação, porém nem sempre são de fácil compreensão pelos interessados ou envolvidos com o mesmo, devido às linguagens e notações técnicas formais utilizadas para captura e representação conceitual de suas informações. Uma Linguagem Cidadã utilizada para dar melhor entendimento ao cidadão pode permitir uma melhor compreensão também para outros papéis envolvidos com o processo. Este artigo apresenta uma proposta de extensão de uso de uma Linguagem Cidadã de modo a buscar melhorar o entendimento de outros perfis envolvidos com os processos organizacionais.*

1. Introdução

Cada organização, seja ela uma entidade governamental, sem fins lucrativos ou uma empresa, precisa gerenciar uma grande quantidade de processos [DUMAS et. al., 2013]. Além de gerenciar processos, as organizações para serem produtivas precisam coordenar o trabalho [PAIM et. al., 2009]. Para tal, empresas ao redor do mundo estão executando iniciativas de Gestão de Processos de Negócios (*Business Process Management* – BPM) com o objetivo de, por exemplo, superar seus competidores e atender às demandas dos órgãos regulamentadores [DUMAS et. al., 2013].

Desde o início do século XXI a intensidade da demanda e a atratividade da gestão de processos por parte das organizações tem aumentado, se mostrando uma prática eficaz na promoção da integração, dinâmica, flexibilização e inovação organizacional [PAIM et. al., 2009].

Modelagem de processos também tem se mostrado fundamental para a compreensão da organização. Observa-se a diminuição das dúvidas sobre a execução de

um processo quando ele está representado através de um modelo [BALDAM et. al., 2014]. Os modelos de processo permitem um registro de informações sobre o negócio, criam um entendimento compartilhado, e se mostram extremamente importantes na solução de problemas e execução de mudanças, dentre outros propósitos [ROSING et. al., 2015].

As notações mais comumente utilizadas na construção dos modelos de processos mostram-se bastante complexas apesar de serem uma opção interessante na redução de problemas de interpretação [AALST, 2003]. Estas são utilizadas no lugar de linguagem natural, que, apesar de parecer de mais fácil entendimento, é passível de problemas semânticos e interpretativos [DEVILLERS, 2011]. Em geral assume-se que os participantes internos à organização são capazes de entender os modelos conceituais, o que nem sempre pode ser verdadeiro [PRILLA E NOLTE, 2012].

No âmbito das empresas públicas, a Lei de Acesso à Informação [LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011] garante ao cidadão acesso à informação considerada pública de forma compreensível e acessível. Dentre os tipos de informação que, por lei, devem ser apresentadas ao cidadão estão os processos, ou seja, a forma como o trabalho é executado dentro de uma organização para gerar informações. Vamos utilizar como exemplo um processo de quebra de requisitos de uma disciplina em uma universidade pública. Este processo deve ser do conhecimento de todos os cidadãos e não somente dos usuários do processo. Pela Lei de Acesso à Informação o cidadão deve ter este acesso a informações descritas numa linguagem que possa ser entendida por ele, chamada pela Lei de "Linguagem Cidadã". Porém a Lei não indica sintaxe ou semântica para esta linguagem, deixando assim a possibilidade desta construção. Trabalhos anteriores mostram que uma Linguagem Cidadã baseada em modelos de processos de negócio pode dar bom entendimento ao cidadão sobre o funcionamento dos processos de uma organização [CARVALHO, et. al., 2016].

Este trabalho sugere o uso de uma Linguagem Cidadã não só pelo cidadão, mas também pelos envolvidos no processo, exemplificando como a mesma poderia ser usada para melhoria do entendimento de outros atores envolvidos com o processo que não tenham conhecimento técnico das linguagens ou notações técnicas.

O artigo se estrutura desta forma: a seção 2 apresenta o conteúdo teórico que o fundamenta a seção 3 apresenta um exemplo de uso de uma Linguagem Cidadã para outros perfis participantes do processo que não o cidadão, a seção 4 concluí o trabalho com as contribuições, limitações deste trabalho e propostas de trabalhos futuros.

2. Fundamentação Teórica

Esta seção objetiva esclarecer termos e conceitos deste artigo: Gestão de Processos de Negócios, modelagem conceitual de processos de negócios e Linguagem Cidadã.

2.1 Gestão de Processos de Negócios (BPM)

BPM é uma disciplina que envolve modelagem, automação, execução, controle, mensuração, e otimização dos fluxos de atividades de negócios em combinações aplicáveis que suportem os fins organizacionais, abrangendo limites organizacionais e sistêmicos e envolvendo funcionários, clientes e parceiros dentro e fora dos limites organizacionais [ROSING et. al., 2015].

Organizações são compostas, dentre outros componentes, por processos de negócios [LAUDON E LAUDON, 2013] e para que eles sejam gerenciados antes devem ser compreendidos [DUBANI et. al., 2010]. Eles envolvem pessoas, comunicações de vários tipos e também mudanças, não sendo apenas junção de software e máquinas. Melhorar a coordenação e integração do trabalho, auxiliar a organização a entender melhor a sua cadeia de valor e prover uma visão sistêmica das atividades da organização [BALDAM et. al., 2014] são exemplos de vantagens da visão processual, que também são o foco deste trabalho. Uma parte essencial da Gestão de Processos de Negócios como conceito de gerência é a modelagem de processos de negócios, requerendo uma abstração dos processos do mundo real objetivando mapeá-los em modelos de processos [SCHREPFER, 2010].

2.2 Modelagem Conceitual de Processos de Negócios

A modelagem de processo de negócio busca prover uma descrição holística de um processo operacional e considera não apenas os problemas de coordenação de tarefas e fluxo de controle, mas também definição de dados e uso no contexto do processo e do ambiente organizacional (estrutura organizacional, recursos, linhas de comando, etc.) na qual irão operar [RUSSELL, et. al., 2016].

As linguagens de modelagem de processos de negócios podem ser divididas em três: formais, conceituais e executáveis [AALST, 2013]. Modelos de processos de negócios podem ser utilizados a fins descritivos, analíticos e executáveis [DEVILLERS, 2011]. Os modelos tratados neste trabalho são conceituais descritivos.

Modelos conceituais possuem um alto nível de abstração, não tendo uma semântica bem definida, não podem ser utilizados para análise e execução direta [AALST, 2013]. Uma de suas funções primárias é representar os processos de forma simples [BALDAM et. al., 2014], garantindo uma visão holística e completa do processo [ROSING et. al., 2015], por exemplo: atividades, comportamentos, recursos, eventos, agentes, etc. [LIN et. al., 2002].

Modelos descritivos tem o propósito de facilitar o entendimento de todos os participantes no processo de negócio e contribuir para alcançar consenso na maneira que o negócio almeja atingir seus objetivos [DEVILLERS, 2011]. Elencam-se vinte casos de uso realísticos para Gestão de Processos de Negócios (BPM), onde onze deles envolvem modelos descritivos [RUSSELL, et. al., 2016], dada sua importância. São exemplos de linguagens e notações técnicas conceituais, entre outras: *Business Process Model and Notation* (BPMN) [OMG, 2011], a *Event-driven Process Chain* (EPC) [AALST, 1999], a *Unified Process Language* (UML) [OMG, 2015].

2.3 Linguagem Cidadã

Modelos de processos de negócio são artefatos centrais para documentação e análise em BPM. Estes modelos, todavia, são normalmente artefatos exclusivos à especialistas da disciplina, sendo de difícil uso pela maioria dos envolvidos [PRILLA E NOLTE, 2012]. Uma Linguagem Cidadã tem como objetivo a melhoria da compreensão dos modelos de processo pelo cidadão. Ela sintetiza os elementos da representação de modelos de processo em modelos conceituais dando a eles mais simplicidade, clareza, legibilidade entre outras características, propiciando melhor entendimento [CARVALHO et. al., 2015]. A Linguagem Cidadã apresentada por Carvalho et. al. (2015) busca mesclar os

benefícios da representação diagramática dos processos de negócios com a facilidade da interpretação da linguagem natural.

Processos de negócios modelados em linguagens e notações técnicas podem ser adaptados para uma Linguagem Cidadã. O cumprimento da legislação ou estratégia organizacional podem ter como requisito a compreensão dos modelos de processos de negócios, e suas informações, não apenas pelos analistas ou especialistas do negócio, mas também por interessados ou envolvidos, externos ou internos.

A Figura 1 apresenta um método que pode ser utilizado para adaptar um modelo conceitual à Linguagem Cidadã proposta por Carvalho et. al. (2015). Depois de realizada a modelagem do processo utilizando uma linguagem ou notação técnica, o modelo é adaptado utilizando um catálogo ou guia de adaptação ao entendimento visando maior efetividade e compatibilização ao o cenário e público-alvo. O Catálogo (características, operacionalizações e mecanismos de implementação) utilizado é uma adaptação do Catálogo de Características de Entendimento de Modelos de Processos de Prestação de Serviços Públicos [ENGIEL, 2012] construída por Carvalho et. al. (2015).

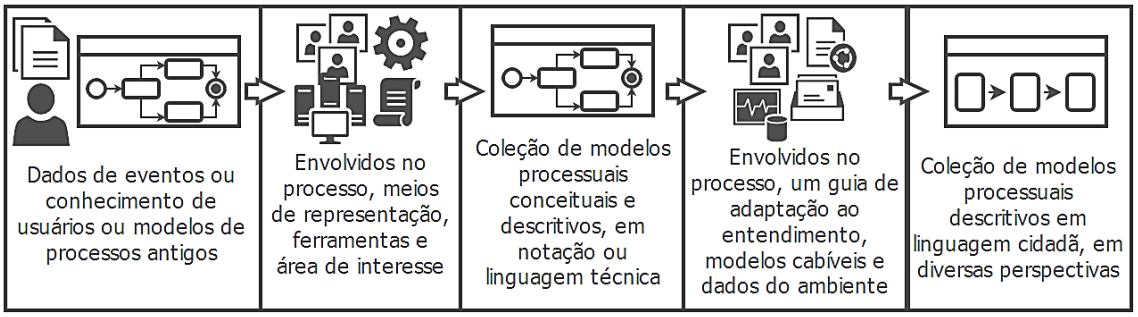


Figura 1: Método de modelagem conceitual e transformação em uma Linguagem Cidadã

Exemplificando, a Figura 2 apresenta um processo em BPMN que quando transformado pelo método da Figura 1, resulta na Figura 3.

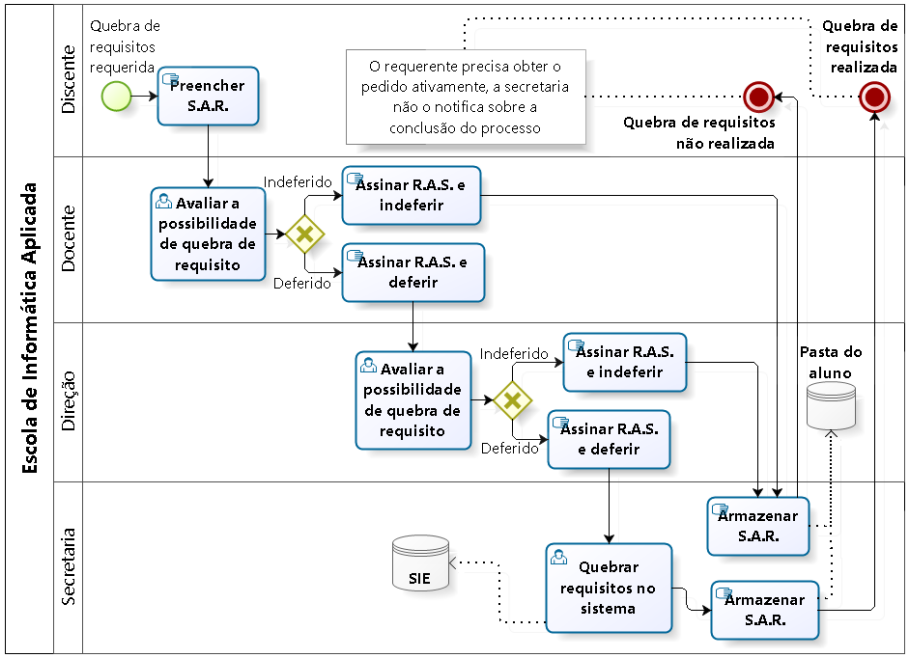


Figura 2: Modelo conceitual descritivo, processo Quebra de Requisito – BPMN

O modelo conceitual apresentado na Figura 2 difunde e compartilha as informações processuais, só que falha em instruir o seu público alvo (cidadão), já que o mesmo não consegue compreender diversos elementos nele representados [CARVALHO et. al., 2015], como por exemplo o elemento "conector lógico" que não expressa por si nenhuma semântica que seja capaz de dizer que sua proposta é uma tomada de decisão [OMG, 2011].

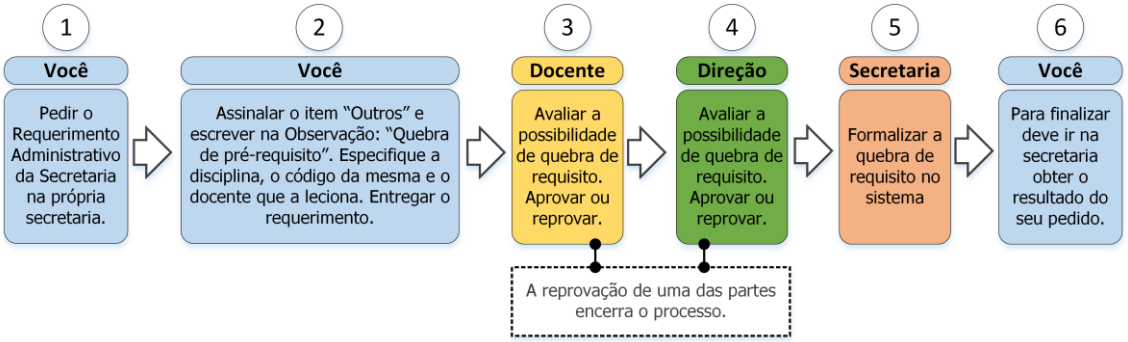


Figura 3: Modelo em Linguagem Cidadã na perspectiva do discente (público alvo)

A Linguagem Cidadã na modelagem, assim como na linguística e no letramento, considera em primeiro lugar a possibilidade de o cidadão ter elementos que garantam a resolução de seus desafios diários no tocante aos textos e discursos, nos diversos espaços sociais e institucionais. Enquanto na análise linguística há uma gama de discursos inerentes aos documentos que impossibilitam seu entendimento [SOUSA E NETO, 2013], na análise gráfica dos modelos conceituais há uma gama de símbolos inerentes à diagramação técnica que seguem esta impossibilidade.

Alguns pontos são importantes de serem destacados quando comparamos a Linguagem Cidadã e a Modelagem Conceitual a partir de linguagens e notações técnicas [CARVALHO et. al., 2015], [CARVALHO et. al., 2016], [ENGIEL, 2012]:

i. Seu foco não está no artefato construído em si, mas sim na efetividade do entendimento dele pelo cidadão;
ii. Não busca prover aos interessados uma visão integral do processo de negócio;
iii. Possui sintaxe e semântica fluída;
iv. Não tem como propósito servir de base para análise do negócio, automação ou implementação de um sistema devido à pouca formalidade de sua semântica;
v. Busca garantir o entendimento pelos interessados nos processos de uma organização, estes fazendo parte dela ou não.
vi. Está alinhada a Lei de Acesso à Informação [LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011];
vii. Pode ser adaptada mais facilmente para corresponder a requisitos de acessibilidade (como braile e melhor interpretação pelos leitores de tela em computadores para deficientes visuais, fontes maiores para idosos, etc.).

3. Aplicando uma Linguagem Cidadã para outros envolvidos no processo

Um modelo incompreensível aos envolvidos conota exclusão dos mesmos na aplicação das iniciativas de processos [NOLTE E PRILLA, 2013]. Para que todos os envolvidos possam de fato participar de forma adequada da execução de um processo é necessário que tenham entendimento dele. Apesar de linguagens e notações técnicas serem

utilizadas para representar modelos internamente em organizações, percebe-se que nem todos os envolvidos tem o mesmo nível de entendimento dos processos [PRILLA E NOLTE, 2012].

Para construção de um exemplo foi aplicado o método apresentado na Figura 1 e utilizado o processo de "Quebra de Requisito" da Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro apresentado na Figura 2. Além do processo para o cidadão apresentado na Figura 3, os modelos nas perspectivas de outros dois atores envolvidos com o processo além do cidadão, podem ser observados nas Figuras 4 (secretaria) e 5 (corpo docente).

Os modelos de processos de negócios em Linguagem Cidadã limitaram-se à perspectiva do recurso discente, então as informações apresentadas focam nele, não havendo excesso de informação e complexidade no mesmo [CARVALHO et. al., 2015]. Os demais envolvidos no processo (docente, direção e secretaria) devem ter, cada um, sua visão específica de modo a entender como o processo funciona para ele como ator principal.

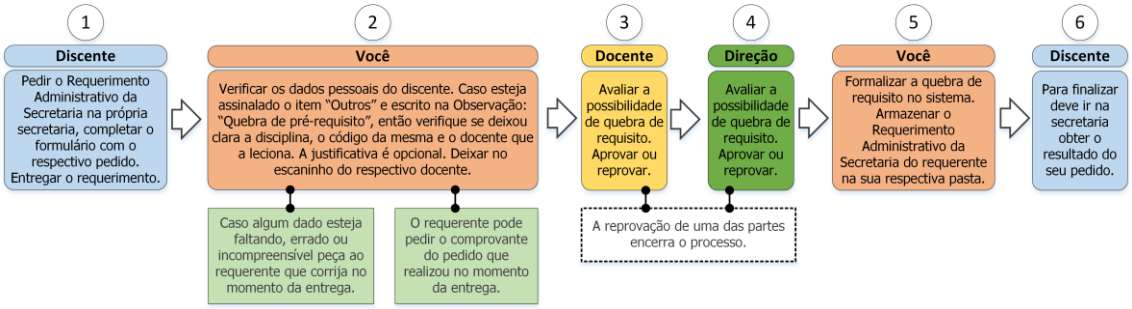


Figura 4: Modelo em Linguagem Cidadã na perspectiva da secretaria

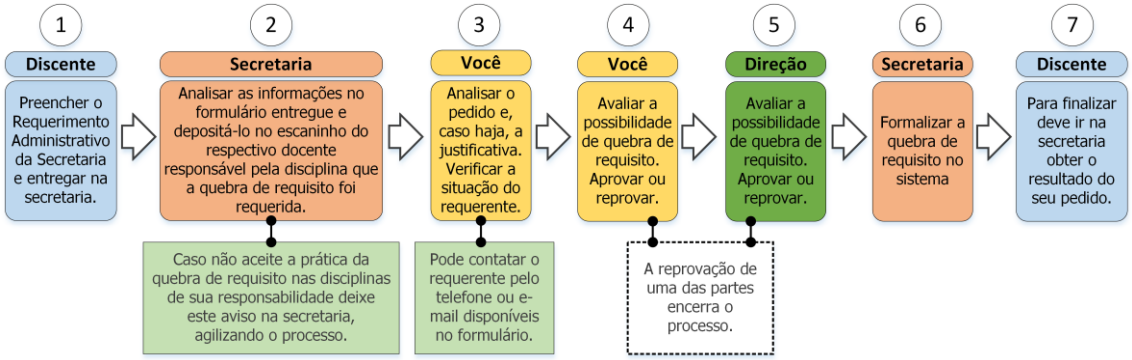


Figura 5: Modelo em Linguagem Cidadã na perspectiva do corpo docente

As Figuras 3, 4 e 5 possuem informações diferenciadas dedicadas aos respectivos envolvidos. O discente (Figura 3) sabe que o processo em algum momento terá tarefas realizadas pelo corpo docente, pela direção e pela secretaria, porém não as conhece integralmente de forma que consiga executá-lo com a visão de outro envolvido. Semelhante para a secretaria na Figura 4 e para o corpo docente na Figura 5.

As Figuras 3, 4 e 5 apresentam modelos diferentes. Cada um deles tem a visão de um envolvido específico, esta visão específica do envolvido apesar de não dar a ele a visão completa do processo e nem a visão de outros envolvidos, facilita o entendimento de suas próprias atividades no processo e também das relações destas com atividades de outros usuários.

Este exemplo mostra que uma Linguagem Cidadã pode ser útil para o caso de existirem na organização envolvidos com o processo, que assim como o cidadão não tem conhecimento de linguagens e notações técnicas que compõe um modelo conceitual.

4. Conclusão

Modelos utilizando uma Linguagem Cidadã possuem informações reduzidas e dedicadas à uma perspectiva particular, pois o excesso de informações dificulta a compreensão de um modelo [ENGIEL, 2012], logo, os mesmos apresentam limitação da representação do domínio e de expressividade em vista de aumentar a simplicidade e compreensibilidade pelo público-alvo. Construir modelos em Linguagem Cidadã na perspectiva de cada um dos envolvidos preserva sua expressividade [RUSSELL, et. al., 2016], pois a soma das informações em todos os modelos em Linguagem Cidadã será equivalente às informações presentes no modelo conceitual descritivo. Não há perda de informação entre a Figura 2 e a soma das informações nas Figuras 3, 4 e 5. Observando os modelos apresentados em uma Linguagem Cidadã neste trabalho pode-se perceber a possibilidade de um melhor entendimento.

Mesmo mantendo a expressividade desta forma ainda há o problema semântico, denotado na precisão e analisabilidade [RUSSELL, et. al., 2016]. Esta Linguagem Cidadã está em um nível de representação entre a linguagem natural e a modelagem conceitual, tendo, portanto, uma sintaxe e semântica mais relaxada, permitindo ainda o desenvolvimento de melhorias na mesma.

Um outro ponto a ser verificado no futuro é a complexidade dos modelos. Modelos mais complexos podem ser mais difíceis de serem adaptados para uma Linguagem Cidadã.

Como trabalho futuro sugere-se a aplicação das perspectivas nos demais modelos da iniciativa, não apenas no modelo do processo “Quebra de Requisito”, mas em outros processos da Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro e a utilização em cenários maiores e mais complexos.

Vê-se também como necessidade futura de avaliação, por exemplo, da melhoria do entendimento desta Linguagem Cidadã por parte dos envolvidos.

5. Referências

Baldam, R., Valle, R., Rozenfeld, H., "Gerenciamento de Processos de Negócios BPM - Uma referência para implantação prática." Campus, 2014.

Carvalho, L. P., Santoro, F., Cappelli, C., “Transparência e entendimento de processos em uma universidade pública”. Em: WTRANS, 2015.

Carvalho, L. P., Santoro, F., Cappelli, C., “Um estudo sobre o entendimento de processos através de modelos com base no público alvo”. Em: II ERSI, 2015.

Carvalho, L. P.; Santoro, F.; Cappelli, C., "Using a citizen language in public process models: the case study of a Brazilian university.". Em: EGOVIS, 2016.

Deviller, M., "Business Process Modeling as a Means to Bridge The Business-IT Divide". Dissertação de mestrado, Universidade Radboud Nijmegen, 2011.

Dubani, Z., Soh, B., Seeling, C., "A Novel Design Framwork for Business Process Modeling in Automotive Industry.". Em: Quinto Simpósio Internacional IEEE em Design, Teste e Aplicações Eletrônicas, Anais eletrônicos, 2010.

Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A., “Fundamentals of Business Process Management”. Springer-Verlag, 2013.

Engiel, P. “Projetando o entendimento de modelos de processos de prestação de serviços públicos. ” Dissertação de mestrado, Universidade Federal do Estado do Rio de Janeiro, 2012.

LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011.

Laudon, K., Laudon, J., "Essentials of Management Information Systems.". Pearson, 2013.

Lin, F.-R., Yang, M.-C., PAI, Y.-H, "A generic structure for business process modeling.". Em: Revista de Gestão de Processos de Negócios, Volume 8, 2002.

Nolte, A., Prilla, M., "Anyone can use models: Potentials, requirements and support for non-expertmodel interaction.". Em: International Journal of e-Collaboration, 2013.

OMG, "Unified Modeling Language (UML) v2.5", <http://www.omg.org/spec/UML/>, 2015.

OMG, “Business Process Model and Notation (BPMN) v2.0”, <http://www.omg.org/spec/BPMN/2.0/>, 2011.

Paim, R.; Cardoso, V.; Caulliraux, H.; Clemente, R., "Gestão de processos: pensar, agir e aprender.". Bookman, 2009.

Prilla, M., Nolte, A., "Integrating Ordinary Users into ProcessManagement: Towards Implementing Bottom-Up, People-Centric BPM.". Em: Lecture Notes in Business Information Processing, 2012.

Rosing, M., Scheer, A., Scheel, H., "The Complete Business Process Handbook - Body of Knowledge from Process Modeling to BPM". Volume 1, Morgan Kaufmann, 2015.

Russel, N., Van der Aalst, W., Hofstede, H., "Workflow Patterns - The Definitive Guide.". The MIT Press, 2016.

Schrepfer, M., “Modeling Guidelines for Business Process Models”. Dissertação de mestrado, Universidade Humboldt de Berlim, 2010.

Sousa, F., Neto, J., "Linguagem cidadã: meio de desburocratização da informação nas instituições brasileiras.". Em: VI Encontro Ibérico EDICIC, 2013.

Van der Aalst, W., "Business Process Management: A Comprehensive Survey.". Hindawi, 2013.

Van der Aalst, W., "Formalization and Verification of Event-driven Process Chains.". Em: Information and Software Technology, Volume 41, 1999.

Van der Aalst, W., Hofstede, A., Weske, M., “Business Process Management: A Survey”. Springer-Verlag, 2003.

W-SAGE: Ferramenta Web para Análise de Dados Geoespaciais

Raul S. Ferreira[1], **Carlos E. Mello**[1,2]

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

[2]Departamento de Sistemas de Informação
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro, Brasil

raulsf@cos.ufrj.br, carlos.mello@ufrj.br

Abstract. *Density estimation is a important statistical method that helps to determine, in a concise way, the probability of a phenomenom about certain types of data. Some kinds of data doesn't have a known distribution but may reveal too much about certain domains, for instance, the geographic data. This work seeks to create an intuitive interface for geographic data analysis together with non geographic data, using a non-parametric density estimator with parallelization techniques in GPU, with the objective to show important perceptions about the observed database. Are shown the first empirical results about the tool W-SAGE.*

Resumo. *Estimar densidades é uma importante técnica estatística, que ajuda a determinar, de forma mais precisa, a probabilidade de um fenômeno sobre determinados tipos de dados. Certos dados não possuem distribuição conhecida a priori mas podem revelar muito sobre um determinado domínio, como é o caso dos dados geográficos. Este trabalho busca construir um sistema veloz com uma interface intuitiva para análise desses dados geográficos ao utilizar um método estimador de densidade não-paramétrico em conjunto com técnicas de paralelização em GPU, com o intuito de disponibilizar percepções importantes sobre a base de dados observada. São mostrados os primeiros resultados empíricos desta ferramenta, entitulada W-SAGE.*

1. Introdução

A informação geográfica tem grande importância em diversas áreas como, marketing, agricultura, meio ambiente, saúde, planejamento urbano entre outros, ajudando na tomada de decisões e estratégias além de agregar valor como um meio de representação visual mais expressiva do que uma representação discreta. Várias ferramentas podem ser criadas para extrair o máximo de informações agregadas à distribuição espacial, informações essas que não poderiam ser extraídas através do modo convencional de análise de dados não espacial. A informação geográfica já existe há centenas de anos (e.g., mapas) e como em vários aspectos do nosso cotidiano esta também foi e vem sendo alterada pela modernidade tecnológica.

Para lidar com esse tipo de informação, surgiram então os sistemas de informações geográficas (SIG). Os SIG são sistemas utilizados para armazenar, analisar, manter e manipular dados geográficos de maneira automatizada [Bolstad 2005]. Os dados geográficos

utilizados pelos SIG podem ser imagens digitalizadas (e.g., fotos de satélite) ou objetos que representam uma geometria no espaço, chamados objetos espaciais. Esses dados são armazenados e gerenciados por bancos de dados espaciais (objetos geométricos espaciais) e estes se encaixam bem dentro do conjunto de tecnologias que compõe este trabalho. Assim, com o auxílio de um banco de dados geográfico, técnicas de extração, tranformação e carregamento de dados, bancos de dados NoSQL(*Not only SQL*) e a utilização de métodos estatísticos, foi construída uma ferramenta de visualização de dados geográficos entitulada W-SAGE (*Web tool for Spatial Analysis of GEographic data*).

Desta forma, este trabalho está organizado em 6 capítulos, onde este é o primeiro. No segundo capítulo são apresentados os trabalhos relacionados às ferramentas de visualização de dados geográficos. No terceiro capítulo são mostradas as motivações na escolha das técnicas e tecnologias utilizadas neste trabalho, enquanto no quarto capítulo, é explicado com maiores detalhes a metodologia que guiou a construção da ferramenta. No quinto capítulo, encontram-se os resultados dos primeiros experimentos utilizando dados reais e no sexto e último capítulo são apresentadas as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

[Mello 2008] faz uso de objetos espaciais em conjunto com técnicas de clusterização com o intuito de tentar mostrar como a acessibilidade na cidade do Rio de Janeiro pode ser melhor modelada, observada e mostra como a disponibilização dessa visualização através de uma representação geográfica torna muito mais clara essa análise para o usuário do que uma representação puramente textual baseada em números e tabelas.

Em [Patrol 2012] podemos ver o trabalho entitulado *Active Missing Person Map*, realizado no Missouri, um estado norte americano, onde foi desenvolvido um sistema de visualização com um viés geográfico ao mostrar em um mapa do território do estado, as quantidades de pessoas desaparecidas separadas por cidade e suas respectivas informações.

[Spina 2016] desenvolveu um trabalho de visualização de dados geográficos voltado para auxiliar o combate ao tráfico humano fazendo uma análise sobre o problema de lavagem de dinheiro. Já em [Zuo et al. 2016] o foco é em desenvolver ferramentas de visualização para auxiliar o descobrimento de padrões geoquímicos na área de geologia, como anomalias e determinadas restrições dentro de uma região.

Outro trabalho recente na área de visualização é descrito em [Guo et al. 2015], onde é proposto um novo *workflow* para tentar lidar com problemas geográficos que possuem a necessidade de computação intensa, usando métodos de paralelização tanto na parte de visualização quanto na parte de processamento, algo parecido com a proposta deste trabalho.

3. Visualização e processamento de dados geográficos

Como pode-se ver, vários trabalhos recentes procuraram lidar com alguma dificuldade no processamento de informações geográficas tanto no lado cliente quanto no lado servidor. Várias técnicas podem ser combinadas para proporcionar *insights* aos especialistas. No caso deste trabalho, iremos agregar em um sistema de visualização geográfico um algoritmo de inferência estatística, rasterização e clusterização de pontos, além de usarmos bancos de dados específicos para tarefas específicas dentro do fluxo da ferramenta.

3.1. Estimando Densidades

Estimar densidades sobre uma população é um trabalho importante e geralmente usamos uma função de densidade para isso. A densidade de uma população pode ser estimada com várias técnicas estatísticas, porém estatisticamente, alguns dados ou populações não possuem estruturas ou parâmetros característicos, no caso, estes dados são conhecidos como não paramétricos. Dados não paramétricos não dependem de dados pertencentes a nenhuma distribuição particular. Tipicamente, o modelo não-paramétrico cresce no sentido de acomodar a complexidade dos dados. Como métodos não paramétricos fazem menos suposições, a aplicabilidade deles é mais larga que os correspondentes métodos paramétricos. Em particular, eles podem ser aplicados em situações em que menos se sabe sobre o problema em questão. Devido a menor dependência de hipóteses, métodos não paramétricos são mais robustos. Exemplo de dado não-paramétrico: distribuição tem a forma normal, tanto a média quanto a variância não foram especificadas.

A função de probabilidade é um conceito fundamental em estatística e existem diversas técnicas que podem ser empregadas para estimar dados não paramétricos e um dos métodos mais conhecidos, é o estimador de densidade de kernel ou KDE(*Kernel Density Estimation*), também conhecido como Janela de Parzen[Duda et al. 2012]. Neste método são utilizadas funções não-lineares como Gaussianas e Sigmóides para se computar a densidade local de cada instância, logo, este trabalho lança mão desta técnica devido à simplicidade e à eficiência perante a literatura para tratar esses tipos de dados.

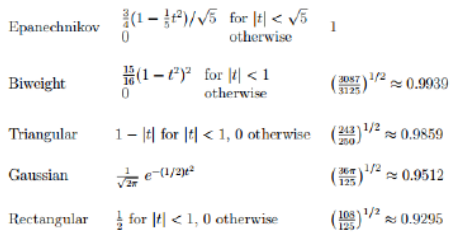
3.2. Kernel Density Estimation

O KDE é uma das técnicas de estimativa de densidade mais comuns e é bastante usada para normalizar e suavizar a distribuição de um determinado conjunto de dados. O KDE pode ser pensado como uma generalização do histograma. Possui duas variações: Univariante, cuja a entrada são dados de uma única dimensão, no caso um vetor; e Multivariante, cuja a natureza dos dados de entrada é de duas ou mais dimensões, usando uma matriz para armazenamento dos dados. É esta versão multivariante que usaremos para estimar as densidades dos pontos geográficos (latitude, longitude).

O algoritmo KDE multivariante é descrito na figura 2 e seu processamento é feito levando-se em consideração cada indivíduo em relação a sua população e a sua complexidade é $O(n^2k)$, pois é implementado como um somatório de um produtório de matrizes e portanto, dependendo do número de dimensões k na entrada, o algoritmo pode-se tornar um tanto quanto lento para apresentar a estimativa de densidade final ou mais conhecido como PDF (*Probability Density Function*), que é calculado para cada indivíduo em relação a sua população.

A partir de um dado número de observações n, calculamos curvas de densidade delas em relação à distância de um valor central, o núcleo, para cada um desses pontos e obtemos a Estimativa de Densidade final somando esses valores. Um kernel é uma função de ponderação padronizada, ou seja, o núcleo determina a suavização do PDF. Esta técnica é amplamente usada em vários algoritmos de aprendizado de máquina, principalmente em SVM (*Support Vector Machines*)[Hearst et al. 1998]. Esta função Kernel precisa ser cuidadosamente escolhida pois pode provocar um super ajustamento (*overfitting*) ou o contrário (*underfitting*) nos valores dos PDFs [Duda et al. 2012][Bishop et al. 2006], no caso deste trabalho será utilizado o Kernel Gaussiano, pois produz uma estimativa mais

suave, porém há outros tipos de funções, como descrito na figura 1.



Epanechnikov	$\frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5}$ for $\lvert t\rvert < \sqrt{5}$, otherwise 0	1
Biweight	$\frac{15}{16}(1-t^2)^2$ for $\lvert t\rvert < 1$, otherwise 0	$\left(\frac{3087}{3125}\right)^{1/2} \approx 0.9939$
Triangular	$1-\lvert t\rvert$ for $\lvert t\rvert < 1$, 0 otherwise	$\left(\frac{243}{250}\right)^{1/2} \approx 0.9859$
Gaussian	$\frac{1}{\sqrt{2\pi}}\, e^{-(1/2)t^2}$	$\left(\frac{36\pi}{125}\right)^{1/2} \approx 0.9512$
Rectangular	$\frac{1}{2}$ for $\lvert t\rvert < 1$, 0 otherwise	$\left(\frac{108}{125}\right)^{1/2} \approx 0.9295$

Figure 1. Tipos de Kernel.

```
for i ← 0 to n do
    soma_kernel ← 0.0
    for j ← 0 to n do
        prod_kernel ← 1.0
        for k ← 0 to xLen do
            prod_kernel *K((x[i][k] − x[j][k])/h)/h
        end
        soma_kernel ← soma_kernel + prod_kernel
    end
    pdf[i] ← soma_kernel / n
end
```

Figure 2. KDE Multivariante.

3.3. Processamento paralelo do KDE em GPU

Devido ao custo de processamento do KDE foi preciso pensar em como melhorar o desempenho do algoritmo. Devido ao fato do algoritmo ser paralelizável em certas partes de seu processamento então foi implementada uma versão que faz uso da paralelização massiva dos passos da aplicação. Utilizamos então, placas gráficas GPU (*Graphics Processor Unit*) através da linguagem CUDA(*Compute Unified Device Architecture*), que devido a sua natureza SIMD (*Single Instruction Multiple Data*) fornece quantidades superiores de processamento e desempenho muito maior do que processadores convencionais [Sanders and Kandrot 2010].

3.4. Sistema de cache, clusterização e rasterização de dados

Para evitar que toda a vez que uma nova consulta fosse feita, uma nova requisição ao banco de dados e um novo processamento do KDE fosse realizado, foi utilizada uma estratégia de cache dos dados de consulta. Para tal, foi utilizado um banco de dados NoSQL, baseado em chave-valor, amplamente utilizado para problemas com alta latência de dados, conhecido como Redis[Han et al. 2011]. Desta forma, este banco funciona como um cache da consulta e assim, melhoramos ainda mais o gargalo de processamento dos PDFs resultantes do KDE e do acesso ao banco de dados geográfico, o que melhora ainda mais a velocidade de resposta da aplicação.

Além disso, devido à grande quantidade de pontos no navegador foi preciso pensar em uma estratégia para diminuir o carregamento da página. A saída foi implementar a clusterização dos pontos por distância, ou seja, pontos que estão muito próximos entre si ou que possuam a mesma coordenada não precisam aparecer individualmente mas apenas um único ponto representando o total de pontos contidos naquela região.

Para isso, foi utilizada a clusterização de pontos no lado cliente da aplicação. Desta forma, ao utilizar a visualização de pontos no mapa da aplicação, caso o usuário esteja em um nível mais distante de zoom então círculos contendo apenas uma contagem total de pontos referentes a uma área aparece e caso o usuário dê o zoom na página então essa distância relativa tenderá a aumentar e consequentemente os pontos, que antes estavam clusterizados, aparecerão em outros *clusters* menores e assim sucessivamente até aparecer o próprio ponto. Dessa forma, evitamos que vários pontos fiquem agrupados em

um espaço pequeno ou até mesmo um em cima do outro, dificultando a visualização e deixando a página desnecessariamente carregada.

Outro ponto importante na implementação foi o uso de rasterização dos pontos, onde convertemos uma imagem vetorial em uma imagem raster (pixels ou pontos) e colocamos em *buffer*. Dessa forma os pontos deixam de ser objetos e caso uma fique sobre a outra a imagem acima tende a sobrescrever o que estiver abaixo dela, evitando que múltiplas imagens empilhadas sejam criadas na tela.

4. W-SAGE: *Web tool for Spatial Analysis of GEographic data*

Foram realizados dois estudos de caso com a utilização do W-SAGE e este foi construído em 2 partes: Lado servidor ou mais comumente chamado de back-end, responsável pelo processamento do KDE e pelo cache das consultas ao banco de dados geográfico; Lado cliente ou simplesmente front-end, responsável por intermediar a interface do usuário com o servidor além de garantir a visualização das consultas em formato de cluster, mapa de calor e gráficos. No primeiro estudo de caso, foi usada uma base de dados contendo a localização das agências do IBGE pelo país. Já no segundo caso de teste, foi utilizada uma base anonimizada referente à alguns dados de matrícula dos alunos da Universidade Federal Rural do Rio de Janeiro (UFRRJ).

4.1. Coleta dos dados

Os dados do IBGE foram coletados no próprio site, nesta base aproveitamos apenas os atributos de identificação numérica e as coordenadas geográficas, totalizando 570 pontos representando as agências do IBGE pelo país. Para o segundo experimento, foram coletados junto à diretoria da UFRRJ, em planilha excel, alguns dados de matrícula de todos os alunos que se matricularam na universidade de 2000 até 2013, totalizando 14027 registros. Os dados fornecidos foram: Cep, situacao da matrícula, status de bolsista, sexo, nascimento, naturalidade, forma de ingresso, periodo real, período cronológico, campus, código do curso, cr acumulado e percentual integralizado. Não houve coleta de nomes ou qualquer tipo de dado sócio-econômico.

4.2. Modelo da aplicação

Conforme pode ser visto na figura 3, o sistema foi desenvolvido com uma arquitetura própria para coleta de dados heterogêneos que podem ser facilmente guardados como documentos no banco de dados NoSQL orientado à documentos, no caso deste trabalho, o banco escolhido foi o MongoDB[Han et al. 2011].

O modelo da aplicação foi desenvolvido pensando-se em uma arquitetura simples de automação, onde num primeiro passo coleta-se os dados de diferentes fontes, colocando-os em um banco de dados orientado à documentos. Em seguida, processos de ETL(*Extract, Transform and Load*) podem ser definidos de acordo com os dados coletados para em seguida serem jogados no banco de dados relacional, a partir daí, a aplicação pode ou não processar uma consulta, baseada no cache salvo pelo banco de dados chave-valor, Redis.

O cache na aplicação foi implementado da seguinte forma: Assim que uma consulta é feita, gera-se um *hash* dos parâmetros desta consulta e salva-se este *hash* como chave e o resultado da consulta como o valor. Caso o usuário faça uma nova consulta, é

feito o mesmo processo descrito e caso a chave seja igual ao que o Redis já possui, este traz o valor sem precisar ir ao banco e caso seja uma nova chave, o Redis guarda esta nova chave e o resultado da consulta como valor, e assim por diante.

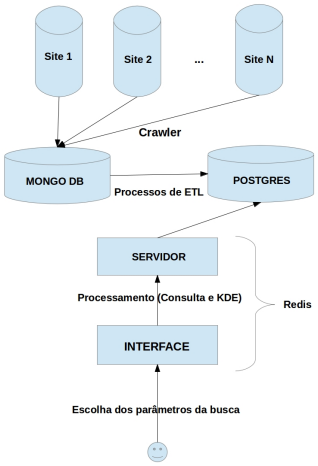


Figure 3. W-SAGE: Fluxo da aplicação

4.3. Visualização dos dados

Depois de enviada a requisição de consulta para o servidor, este recupera os dados e em seguida processa o KDE que por sua vez entrega o resultado, no caso, um vetor de pdf, para que seja criado o mapa de calor onde atribui-se aos pesos dos pontos os valores dos PDFs, gerando no mapa, colorações que variam do azul (menor peso) ao vermelho (maior peso). Quanto maior o peso maior a probabilidade de ocorrer o fenômeno estudado. Os pontos então são processados através de clusterização e rasterização e colocados no mapa juntamente com o mapa de calor e os demais gráficos.

5. Resultados

Os resultados apresentados estão balizados em três premissas:

1. Verossimilhança do resultado visual do estimador de densidade processado pelo sistema em relação ao mesmo algoritmo fornecido em um software comercial;
2. Velocidade de processamento do KDE em um software comercial comparado ao W-SAGE;
3. Facilidade de visualização e suporte a tomada de decisão através da ferramenta.

Como parâmetro para comparação, foi usado o já consagrado software comercial Matlab[Guide 1998] para rodar o KDE gaussiano com as mesmas coordenadas usadas na aplicação e depois seus tempos de processamento foram comparados. Em seguida, a imagem gerada pelo Matlab foi comparada com o mapa de calor gerado pela aplicação, para assim podermos comparar a distribuição gerada pelos dois métodos.

A figura 4 mostra o resultado do KDE gerado nos dados do IBGE pelo matlab e a figura 5 mostra o resultado do KDE gerado nos dados do IBGE pelo W-SAGE. Na figura 6 e na figura 7 são mostrados o KDE no MATLAB e no W-SAGE respectivamente, enquanto na figura 8 uma visão mais completa do mesmo resultado no W-SAGE, já com

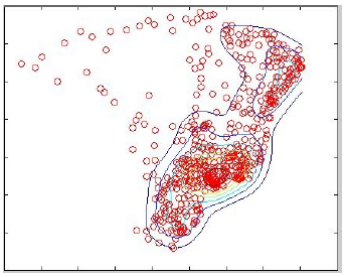


Figure 4. KDE gerado nos dados do IBGE pelo MATLAB.

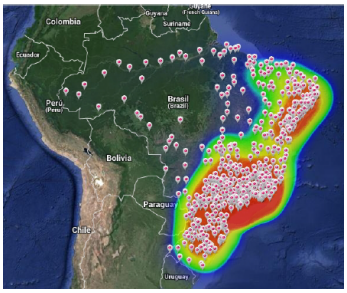


Figure 5. KDE gerado nos dados do IBGE pelo W-SAGE.

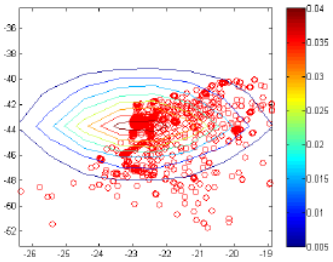


Figure 6. KDE gerado nos dados da UFRRJ pelo MATLAB.

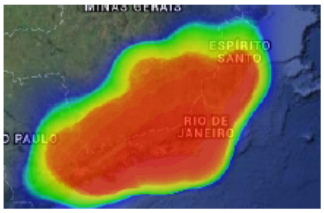


Figure 7. KDE gerado nos dados da UFRRJ pelo W-SAGE.

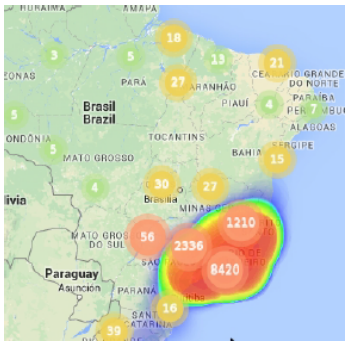


Figure 8. W-SAGE: Resultado de uma consulta.

os pontos rasterizados, clusterizados e com mapa de calor provindo dos PDFs calculados. A velocidade de processamento foi calculado com detalhes em [Ferreira et al. 2014] e um resumo pode ser visto na tabela 1 onde o ganho do KDE paralelizado com GPU foi de 3024% sobre a versão sequencial e 6% superior à versão paralelizada do MATLAB.

Table 1. Tabela comparativa

Tabela 1. Tempos medidos.

KDE Sequencial	KDE Paralelizado Matlab	KDE c/ CUDA	KDE c/ CUDA otimizado	Speed-Up (GPU x Serial)	Speed-Up (GPU x Matlab)
31.088s	6.355s	1.680s	1.028s	30,241	6,181

6. Conclusão

Através deste trabalho conseguimos integrar um método estimador de densidades, com resultados parecidos com o de softwares reconhecidos pelo mercado, à uma aplicação web de análise de dados geográficos. A ferramenta se mostra escalável para o carregamento de milhares de pontos no browser se for necessário, ao paralelizar uma parte do algoritmo usando GPU e ao utilizar técnicas de clusterização e rasterização de coordenadas geográficas. Além disso, foi utilizada uma estratégia para lidar com dados heterogêneos além da utilização de *caching* evitando reprocessamento de consultas. O sistema proporciona assim, a análise dos dados geográficos, não só limitada a estes, de forma intuitiva, com qualidade e com boa velocidade de resposta.

6.1. Trabalhos Futuros

Como trabalho futuro pretendemos utilizar bancos de dados distribuídos e uma arquitetura assíncrona preparada para receber muitas requisições (Node.js), para tentarmos escalar o sistema para *datasets* com milhões de pontos. Outro ponto importante é a questão da substituição do KDE tradicional por um modelo mais rápido do algoritmo, recentemente proposto por [O'Brien et al. 2016] onde a complexidade do KDE, que é quadrática, foi reduzida para linear.

References

Bishop, C. M. et al. (2006). Pattern recognition and machine learning, vol. 1springer. *New York*, (4):12.

Bolstad, P. (2005). *GIS Fundamentals: A First Text on Geographic Information Systems.* Eider Press.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Ferreira, R. S., Valenzuela, J. E. H., and Zamith, M. P. (2014). Paralelização do algoritmo de método de estimação não-paramétrico por núcleo estimador multivariado (kde) utilizando gpu/cuda. In *II Reunião Anual de Iniciação Científica da UFRRJ*.

Guide, M. U. (1998). The mathworks. *Inc., Natick, MA*, 5:333.

Guo, M., Guan, Q., Xie, Z., Wu, L., Luo, X., and Huang, Y. (2015). A spatially adaptive decomposition approach for parallel vector data visualization of polylines and polygons. *International Journal of Geographical Information Science*, 29(8):1419–1440.

Han, J., Haihong, E., Le, G., and Du, J. (2011). Survey on nosql database. In *Pervasive computing and applications (ICPCA), 2011 6th international conference on*, pages 363–366. IEEE.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Mello, C. E. R. (2008). *Agrupamento de regiões: Uma abordagem utilizando acessibilidade*. PhD thesis, UNIVERSIDADE FEDERAL DO RIO DE JANEIRO.

O'Brien, T. A., Kashinath, K., Cavanaugh, N. R., Collins, W. D., and O'Brien, J. P. (2016). A fast and objective multidimensional kernel density estimation method: fastkde. *Computational Statistics & Data Analysis*, 101:148–160.

Patrol, M. (2012). Active missing person map.

Sanders, J. and Kandrot, E. (2010). *CUDA by example: an introduction to general-purpose GPU programming*. Addison-Wesley Professional.

Spina, M. (2016). *New techniques for combatting human trafficking; specifically through the analysis of anti-money laundering and geographic data visualization technology*. PhD thesis, UTICA COLLEGE.

Zuo, R., Carranza, E. J. M., and Wang, J. (2016). Spatial analysis and visualization of exploration geochemical data. *Earth-Science Reviews*, 158:9–18.

Usando Workflows Datacêntricos Para Analisar Tweets Sobre o *Aedes aegypit*

Fillipe Dornelas[1,3], Sérgio Manuel Serra da Cruz[1,2]

[1] Departamento de Matemática –Universidade Federal Rural do Rio de Janeiro (UFRRJ)
[2] Programa PET Sistemas de Informação (PET-SI/UFRRJ)
BR 465, KM7 – UFRRJ – Seropédica – RJ – Brasil

[3] IBM Research do Brasil
Avenida Pasteur, 138 - Urca - Rio de Janeiro – RJ - Brasil

fillipes@br.ibm.com, serra@pet-si.ufrrj.br

Abstract. *Analyzing user messages in social media can offer different point of view of a given society, including public health issues. This work presents a strategy based on data-centric workflows able to collect, prepare and analyze large amounts of tweets evaluating the impact of the messages about the Aedes aegypti in the Brazilian public health scenario. Static and temporal analysis were performed by workflows enacted in IBM Bluemix platform which has been shown as stable and scalable platform.*

Resumo. *A análise de mensagens de redes sociais pode oferecer diferentes perspectivas sobre como as populações se relacionam, incluindo áreas da saúde pública. Este trabalho apresenta um estudo inicial baseado no uso de workflows do tipo datacêntricos executados em nuvens de computadores capazes de coletar e preparar e analisar tweets, avaliando o impacto das postagens acerca do mosquito Aedes aegypti no cenário de saúde pública brasileira. As análises ora apresentadas são de natureza estáticas e temporais e foram integralmente realizadas na plataforma IBM Bluemix.*

1. Introdução

Compreender com profundidade comportamentos e assuntos relacionados ao espalhamento dos *tweets* sobre a saúde pública ainda é um grande desafio em aberto na computação. Adicionalmente, se considerarmos as enormes quantidades de dados manipulados na área da saúde pública, a gravidade e a abrangência e a gravidade das doenças transmitidas pelo mosquito *Aedes aegypti* no Brasil, este problema se torna ainda mais crítico.

Vários estudos têm como objetivo avaliar o espalhamento das mensagens em redes sociais sobre eventos sociais, catástrofes e epidemias (Sprenger et al, 2013), (Dalmonte et al, 2014) e (Santos et al, 2015). O microblog Twitter é uma das redes sociais mais utilizadas no mundo e o Brasil ocupa a segunda posição entre os países com maior número de usuários. A rede emergiu como um dos meios de propagação mais profícuos de disseminação de informações (Chew e Eysenbach, 2010 e Kwak et al,

2010). Seu limite de postagem de poucos caracteres é um facilitador para que os usuários, agências governamentais ou do terceiro setor realizem postagens de forma rápida e sucinta e que se tornam uma fonte importante de alertas de situações de emergências.

Nos últimos anos a disseminação do mosquito *Aedes aegypti* tem alcançados proporções alarmantes no Brasil e no mundo, favorecendo a disseminação de doenças virais até pouco tempo negligenciadas (por exemplo, Zika e Chikungunya). Neste trabalho apresentaremos um estudo baseado em workflows datacêntricos executados em nuvem de computadores que permitem analisar grandes volumes de mensagens semi-estruturadas relacionadas com as postagens relacionadas ao tema "*Aedes aegypti"*, utilizamos dados do Twitter coletados por um período de seis meses em todo o Brasil.

Diferentemente dos trabalhos relacionados, concebemos uma estratégia baseada em workflows datacêntricos em ambiente de nuvem do tipo PaaS cuja composição envolve atividades que variam desde a automação da coleta dos tweets, processamento e posterior classificação/análise/visualização e verificação da disseminação das mensagens. A estratégia proposta foi materializada plataforma Bluemix (Kim et al, 2016), ela permitiu analisar questões relacionadas com o espalhamento de tweets sobre *Aedes aegypti*. Para avaliar a abordagem, propomos um conjunto de questões (Q1, Q2, Q3 e Q4) para investigar o espalhamento e testar a viabilidade da abordagem.

Este trabalho está organizado da seguinte forma. Na Seção 2 apresentamos uma visão geral da literatura relacionada sobre estudos de comportamento de usuários no Twitter em relação as doenças transmitidas pelo *Aedes aegypti*. Na Seção 3 descrevemos os materiais e a metodologia utilizada nos workflows centrados em dados, além da caracterização do dataset e as análises realizadas, Na Seção 4, discutimos os principais resultados obtidos. Finalmente, na Seção 5 apresentamos as conclusões, limitações e alguns direcionamentos para trabalhos futuros.

2. Trabalhos Relacionados

O Twitter tem sido usado em diversos contextos, possui canais de cidadania, saúde e emergências sociais que têm despertado grande importância no cenário de análise de dados sociais. Um dos usos que mais vem despertando atenção diz respeito às questões ligadas à saúde pública, em especial aquelas relacionados com as doenças transmitidas por vírus (H1N1, SARS, Dengue, Zika, Malária, entre outros) que podem atingir grandes contingentes populacionais (Van Hilten et al, 2016).

Antunes et al. (2014) usaram os tweets com a ocorrência do termo "dengue" para inferir quais os períodos onde mais se comentava sobre este assunto e onde mais se encontravam registros de casos da doença em uma determinada região analisada. Toriumi et al. (2013) usaram os tweets para elaborar mapas de projeção e abrangência de um determinado assunto, os autores desenvolveram uma aplicação que exibe mapas e informações sobre a provável infestação dos mosquitos *Aedes aegypti* no município de Cuiabá, no Mato Grosso. A partir da seleção e análise desses dados, os autores foram capazes de desenvolver uma ferramenta de fácil visualização e entendimento sobre possíveis infestações e disseminações das doenças virais.

Além dos estudos sobre a disseminação de epidemias, o Twitter também é largamente utilizado em desastres naturais. Por exemplo, existem trabalhos construídos com a perspectiva de analisar como a informação se propagada durante a ocorrência de desastres naturais (Toriumi et al, 2013 e Thapa et al, 2016). O primeiro autor utilizou tweets para estudar como se comportava o compartilhamento das informações durante o terremoto no Leste do Japão de 2013. O segundo analisou o espalhamento de dados das redes sociais Twitter e Flicker relacionadas ao terremoto no Nepal de 2015. Apesar de serem trabalhos independentes, os autores concluíram que os usuários compartilharam tweets colaborativamente para disseminar as informações que consideram importantes acerca do desastre e também diminuíram o compartilhamento de informações não emergenciais para evitar interromper os fluxos das informações críticas.

Como relação a geolocalização dos tweets, verifica-se que grande parte destes não são localizados por opção própria de usuários ou por questões de privacidade; a maioria evita informar suas reais localizações. Segundo Leetaru et al., (2013) apenas 2% das mensagens são geolocalizadas. Com vistas a preencher essa lacuna, Davis Jr. et al., (2011) usaram dados de tweets não geolocalizados e de informações de relacionamentos entre os usuários do Twitter para enriquecer a tentativa de inferir a localização desses tweets a partir da técnica de validação cruzada de informações.

Até o momento, existem alguns trabalhos na literatura que associem o problema de extração e analise de tweets com uso de workflows científicos. Um workflow científico pode ser definido como sendo especificação formal de um processo científico que representa o encadeamento de fluxos de atividades e dados a serem conduzidas em um determinado experimento (Deelman et al, 2009.). Eles são executados por sistemas gerenciadores de workflows científicos (SGWfC) que fornecem o ferramental necessário para definir, modificar, gerenciar, executar e monitorar os workflows científicos. Os workflows do tipo datacêntricos seguem a mesma lógica dos workflows científicos tradicionais, são centrados em grandes volumes de dados complexos e podem ser executados por SGWfC ou não.

Um SGWfC é um sistema computacional que executa aplicações científicas compostas por atividades cuja ordem de execução é definida por uma representação digital da lógica do workflow científico (Goderis et al, 2006). Atualmente, existem dezenas de SGWfC (Kepler, VisTrails, Pegasus, Taverna, Panda, Galaxy, Swift, Knime, entre outros) (Deelman et al, 2009). Os SGWfC são produtos de diferentes motivações de desenvolvimento, públicos-alvo específicos e decisões técnicas particulares a cada projeto, o que faz com que suas funcionalidades se diferenciem consideravelmente um do outro e que representem diferentes aspectos relacionados à execução e à modelagem de workflows científicos.

Faz necessário ressaltar que até o momento da escrita deste trabalho não foram localizados na literatura SGWfC especificamente concebidos para modelar problemas comuns à área de análise de dados de redes sociais. Por esse motivo, investigamos uso de novas ferramentas de *data analytics* tais como plataforma Bluemix da IBM para modelar e executar os workflows datacêntricos.

O Bluemix da IBM é uma plataforma de serviços de nuvem (PaaS) elástica e escalável baseada no projeto de código aberto Cloud Foundry (2016). Ela permite criar, implementar e gerenciar aplicativos na nuvem com baixo esforço de programação. O

Bluemix é uma plataforma comercial que não foi concebida para atuar ou incorporar as funcionalidades de um SGWfC, porém ele oferece um ecossistema aplicativos, componentes e serviços em tempo de execução que permitem que um pesquisador encadeie atividades computacionais de modo análogo a um workflow científico. O encadeamento das atividades se dá por intermédio de editores de workflows (utilizamos o editor Node-RED (2016)). O Node-RED é editor de workflows multiplataforma que possui interfaces ricas baseadas em Javascript e Node.js e que permite ao pesquisador modelar, e monitorar a execução dos workflows datacêntricos que analisam os dados semiestruturados de oriundos de redes sociais.

Diferentemente dos trabalhos principais relacionados na literatura, neste trabalho propomos a adoção do paradigma dos workflows datacêntricos em ambientes elásticos de *data analytics* para analisar os tweets relacionados à disseminação do *Aedes Aegypti.* As análises dos tweets sobre o tema serão realizadas por workflows desenvolvidos e executados em uma plataforma de serviços de computação em nuvem.

3. Materiais e Métodos

Esta seção descreve os materiais, métodos e etapas propostas para a extração, análise, processamento e visualização dos dados do Twitter.

3.1 Materiais

Nosso estudo se considerou apenas o termo "Aedes aegypti", não foram consideradas variações termo. A coleta dos dados considerou os todos tweets postados por usuários de todo o mundo no período que variou entre junho de 2014 até junho de 2016. Durante esse intervalo, coletamos automaticamente um total de 44.467 tweets.

Utilizamos a plataforma Bluemix e as ferramentas de Data&Analytics disponíveis no catálogo de serviços da plataforma para o desenvolvimento e execução dos workflows. Dentre as principais ferramentas utilizadas destacamos: API de extração e recuperação de dados do Twitter. O repositório de dados utilizado foi o dashDB. O dashDB oferece serviços de banco de dados SQL totalmente gerenciado para cargas de trabalho transacionais, ele foi utilizado como área de armazenamento temporário dos dados (consumidos e produzidos) pelo workflow. Além disso, utilizamos o Bluemix para executar as análises estatísticas sobre os tweets, provendo resultados textuais e visualizações gráficas dos resultados produzidos pelos workflows. Para a análise de sentimentos invocamos os recursos de computação cognitiva do *Watson Analytics services* (WATSON, 2016). O Bluemix e o dashDB foram instanciados em máquinas virtuais com 64GB de memória RAM com 20GB de espaço de armazenamento oferecidas pelo serviço de virtualização OpenStack.

3.2 Métodos

Neste trabalho propomos uma abordagem metodológica baseada em quatro fases para analisar os tweets. A representação gráfica das fases está ilustrada na Figura 1, elas foram fundamentais para a modelagem do workflow datacêntrico.

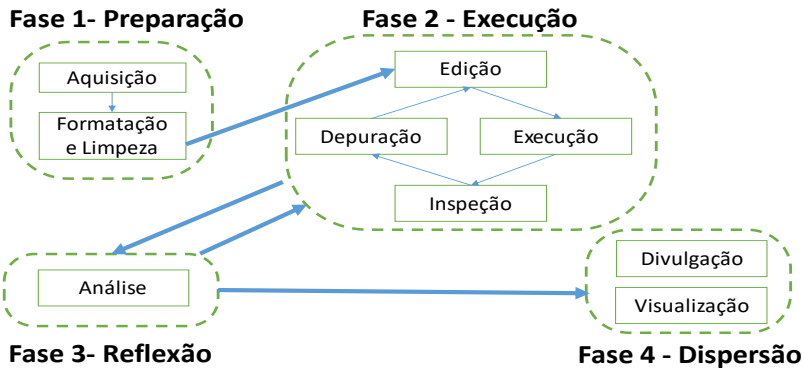


Figura 1. Representação simplificada e conceitual das fases de um workflow datacêntrico para análise de tweets.

A primeira fase (denominada preparação de dados) é executada antes de qualquer tipo de processamento analítico, nela ocorrem a aquisição dos tweets e a preparação ou formatação/limpeza dos dados para serem analisados.

A segunda fase (denominada execução) é o elemento central no workflow datacêntrico. Nela, ocorrem a edição/codificação/encadeamento/execução dos scripts dos workflows. Além disso, ocorrem as análises parciais dos resultados intermediários do experimento, bem como a depuração dos scripts. Essa fase pode ser encarada com um laço, onde o pesquisador interage com a plataforma, realiza múltiplas execuções do workflow com parâmetros distintos para explorar as hipóteses do modelo computacional.

A terceira fase (denominada reflexão) é a eminentemente analítica (ou pós-execução) no processo de exploração dos dados. Comumente, o pesquisador oscila entre as fases de reflexão e execução até a finalização do seu experimento. Nesta fase ele analisa os resultados, inspeciona arquivos, faz anotações e comparações entre as múltiplas execuções do workflow.

Por fim, a quarta fase (denominada dispersão) diz respeito a divulgação, visualização ou compartilhamento dos resultados consolidados obtidos na pesquisa. Nesta fase, ocorrem a publicações dos dados e resultados bem como dos workflows subjacentes.

3.3 Questões de pesquisa

Neste estudo se buscou investigar um pequeno conjunto de questões (Q1, Q2, Q3 e Q4) para testar a viabilidade da abordagem baseada em workflows datacêntricos e também verificar o espalhamento de tweets.

As questões de pesquisa são experimentos baseados no workflow (Figura 2). Q1: Qual foi a contribuição dos tweets em termos de quantos foram os usuários mais influenciadores? Q2: Quais os períodos de maior postagem de tweets sobre o termo “Aedes aegypti”? Q3: Quais as hashtags foram mais postadas no período avaliado? Q4: Qual a predominância dos sentimentos dos tweets?

3.4 Representação conceitual do workflow

Para analisar os dados e verificar a abrangência dos tweets, desenvolvemos um workflow datacêntrico baseados nas quatro fases apresentadas na subseção 3.1. A figura 2 ilustra uma representação conceitual simplificada do workflow baseado nas fases considerando os recursos utilizados para a execução dos experimentos.

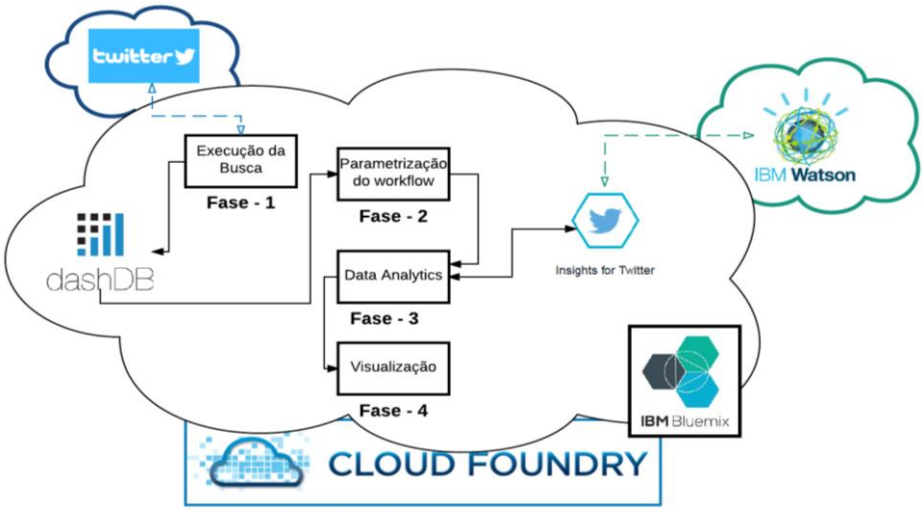


Figura 2. Representação conceitual de um workflow datacêntrico no Bluemix.

4. Provas de Conceito

Com o intuito de avaliar as funcionalidades do workflow foi realizada uma extração de tweets entre os meses supracitados na subseção 3.2. Foram executados experimentos como provas de conceito com o workflow parametrizável. Os experimentos buscavam responder as questões Q1, Q2, Q3 e Q4.

Para responder a Q1, realizamos a execução do workflow que produziu uma simples avaliação do tipo estatística. Dentre todos os tweets da base experimental, verificou-se que existiam apenas 25.370 usuários influenciadores que postaram tweets com o termo avaliado. A abrangência desses alcançaram 255.465.058 milhões seguidores. Como decorrência de Q1, refinamos as análises dos tweets da base experimental e verificamos que apenas 1,83% (818 mensagens) possuíam informações de geolocalização.

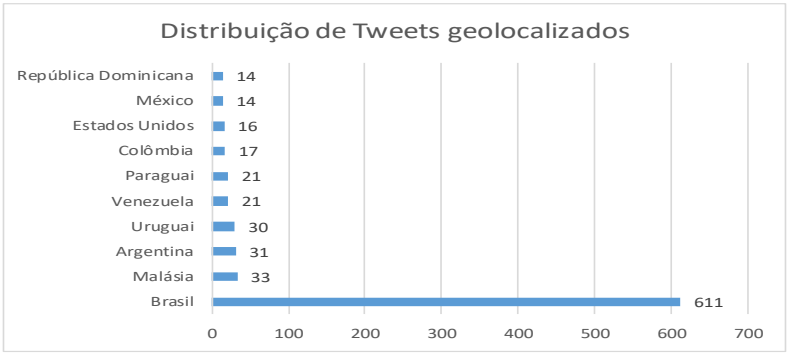


Figura 3. Distribuição de tweets por pais origem.

A Figura 3 apresenta a distribuição dos tweets geolocalizados avaliados pelo workflow. Também se verificou que 43.649 (Tweets e Retweets) não são geolocalizados, sendo que apenas 611 são confirmados do Brasil. Os resultados estão alinhados com as estimativas de apenas 2% de geolocalização de tweets apresentada por (Leetaru et al, 2011).

Para responder Q2, o workflow foi configurado para avaliar a frequência de mensagens. Obtivemos os resultados apresentadas da Figura 4.

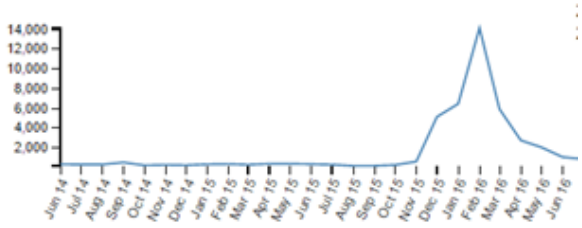


Figura 4. Distribuição temporal de tweets.

Ressaltamos que a questão Q2 difere de Q1, a primeira apresenta apenas resultados estatísticos. Q2 analisa as mensagens em função da sua distribuição temporal e representa os quantitativos de tweets ao longo do período de tempo do estudo. Verificou-se que ocorreu um aumento subido de mensagens sobre o tema nos períodos de novembro de 2015 até março de 2016. Estes períodos correspondem aos meses de verão no Brasil onde ocorre um aumento natural dos casos das doenças transmitidas pelos mosquitos. Além disso, verificou-se que o período se alinhou com a intensa campanha governamental de combate ao mosquito.

Para responder Q3, o workflow foi parametrizado para analisar a frequência e os períodos de postagens dos usuários sobre o tema. Os resultados gerados pelo workflow tiveram como saída a Tabela 1.

Tabela 1. Resumo do quantitativo das hashtags postadas no período avaliado.

Hashtag	Número total de ocorrências
#Zika	374
#ZikaZero	20
#G1	18
#Dengue	13
#CombateAedes	9
Outras	23.783

Por fim, para responder Q4, o workflow foi parametrizado para analisar os sentimentos das mensagens utilizando os algoritmos disponibilizados pelo *Insights for Twitter* que se apoiam os recursos de computação cognitiva oferecidos pelo IBM Watson Analytics Services (WATSON, 2016). Do total de mensagens originais, 42.697 não possui informações de sentimentos. Apenas 3,98% possuíam indicações, sendo 863 identificados como positivas e 575 com sentimentos negativos.

5. Conclusão

Neste trabalho desenvolvemos uma estratégia computacional baseada em workflows datacêntricos apoiados por uma plataforma PaaS comercial de *data analytics* para analisar o espalhamento de tweets relacionados ao tema "Aedes aegypti".

Verificou-se que a plataforma ainda é pouco difundida na comunidade científica, porém ela ofereceu um amplo suporte para o desenvolvimento do workflow e condução dos experimentos. Ela permitiu responder as questões Q1, Q2, Q3 e Q4 com agilidade. Destacamos que, apesar de não ser o foco deste trabalho avaliar o desempenho do workflow datacêntrico, ele produziu os resultados em tempo muito curto, aproximadamente três minutos para todas as execuções.

A abordagem baseada em workflow datacêntrico no Bluemix permitiu que se verificasse que o espalhamento dos tweets avaliados. Observou-se que os períodos de maior número de postagens coincidem com os momentos de maior enfoque do tema nas mídias (rádio, TV e Internet) e nas campanhas publicitárias que alertavam sobre os perigos e formas de prevenção das doenças relacionadas ao mosquito *Aedes aegypti*.

Como limitações encontramos dificuldades ao analisar como os tweets não geolocalizados. Como trabalhos futuros existem diversas possibilidades oferecidas pela plataforma e que por questões de escopo não foram exploradas neste trabalho, como por exemplo aprofundar as análises de sentimentos dos tweets sobre o tema e produzir visualizações dos dados.

Agradecimentos

Agradecemos ao FNDE e ao MEC/SeSU pelo financiamento concedido ao programa PET SI/UFRRJ e a IBM Research do Brasil pelo acesso gratuito aos seus recursos computacionais e a plataforma Bluemix.

Referências

Antunes M. N, et al. 2014. "Monitoramento de informação em mídias sociais: o e-Monitor Dengue", In: TransInformação, Campinas, 26(1):9-18, jan./abr., Brasil.

Cloud Foundry, 2016. https://www.cloudfoundry.org/

Chew C, Eysenbach G. 2010. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One. 2010 Nov 29;5(11):e14118.

Dalmonte, E. 2014. Novos cenários comunicacionais no contexto das mídias interativas: o espalhamento midiático. *Revista Famecos.* DOI: http://dx.doi.org/10.15448/1980-3729.2015.2.19729.

DashDB. 2016. http://www.ibm.com/analytics/us/en/technology/cloud-data-services/dashdb/

Davis Jr, C. A. et al 2011. Inferring the Location of Twitter Messages Based on User Relationships", In: Transactions in GIS, Blackwell Publishing Ltd.

Deelman, E et al. 2009 Workflows and e-Science: An overview of workflow system features and capabilities. Future Generation Computer Systems 25 (5), 528-540.

Goderis, A., Li, P., e Goble, C. 2006. Workflow discovery: the problem, a case study from e-science and a graph-based solution. In Int. Conf. on Web Services (ICWS), pp. 312–319.

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is twitter, a Social Network or a News Media? In Proceedings of the 19th Int Conf. on World Wide Web, pp. 591–600.

Node-Red. 2016. http://nodered.org/.

Santos, H.S et al. 2015. Uma Visão do Mercado Brasileiro de Ações a partir de Dados do Twitter, In: IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015), Brasil.

Thapa, L. 2016. Spatial-Temporal Analysis Of Social Media Data Related To Nepal Earthquake 2015. XXIII ISPRS Congress, July 2016, Prague, Czech Republic, pp. 567-571.

Toriumi, F. et al., 2013. Information Sharing on Twitter during the 2011 Catastrophic Earthquake. In: Proc. 22nd Int'l Conf. on World Wide Web, pp. 1025–1028.

Van Hilten, L. G. 2016. Debunking Zika virus pseudoscience: we need to respond fast, say researchers https://www.elsevier.com/connect/debunking-zika-virus-pseudoscience-we-need-to-respond-fast-say-researchers.

Watson Analytics, 2016. http://www-03.ibm.com/software/products/en/watson-analytics.

Spotify em Foco: Um Estudo de Caso sobre Sistemas para a Terceira Plataforma Computacional

Aíquis Rodrigues[1], Cesar Guimarães[1], José Viterbo[1], Clodis Boscarioli[2]

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

[2] Colegiado de Ciência da Computação
Universidade Estadual do Oeste do Paraná (UNIOESTE)
Cascavel – PR – Brasil

aiquisrg@gmail.com, cesar.portela@hotmail.com, viterbo@ic.uff.br, clodis.boscarioli@unioeste.br

Resumo. *Este trabalho apresenta uma discussão sobre a evolução histórica da computação chegando ao conceito de Terceira Plataforma Computacional. Definimos e discorremos sobre os pilares que compõem a Terceira Plataforma, que são computação em nuvem, big data, web social e mobilidade. Finalmente, através de um estudo de caso abordando uma empresa inserida nesse novo paradigma, o Spotify, identificamos e descrevemos os aspectos relevantes do emprego dessas tecnologias na prática.*

1. Introdução

Quando as primeiras máquinas computacionais foram concebidas durante a Segunda Guerra Mundial, não se imaginava o quanto seu potencial poderia crescer com a evolução tecnológica. Com o passar dos anos foram ficando menores, mais potentes, e passaram a ter usos diversificados, saindo do ambiente militar às empresas e indústrias até chegarem aos lares e cotidiano de todos os cidadãos.

No início havia os gigantescos e primeiros computadores baseados em tubos de vácuo e pouca capacidade de armazenamento de dados. Nos anos 60 os primeiros computadores começaram a ser comercializados. Já em 1964 os tubos a vácuo começaram a ser deixados de lado e a tecnologia do momento eram os transistores.

Uma das características mais fortes dos Mainframes é a grande capacidade computacional, provendo processamento de uma grande quantidade de dados. Nesse período também se iniciaram os processamentos a base de transações, que é um modo de processamento onde existe interação com o usuário, que faz uma requisição e recebe uma saída processada. Para que isso acontecesse fez-se necessário o uso dos chamados Terminais Burros (dumb terminal). Como o nome sugere, este é um dispositivo que não possuía capacidade de processamento e depende totalmente de outro computador.

As décadas de 70 e 80 foram de mudança nos paradigmas computacionais, com a evolução dos mainframes para os computadores pessoais, com os microprocessadores feitos pela Intel, com uma unidade central de processamento em tamanho muito reduzido. Essa inovação foi um facilitador no desenvolvimento de dispositivos menores,

como minicomputadores, monitores e impressoras. Após a consolidação do uso dos microcomputadores pela sociedade, a próxima grande revolução aconteceu no âmbito da comunicação com a evolução e popularização da Internet (e pela World Wide Web – WWW) em meados dos anos 90 e início dos anos 2000. A transição da primeira para a segunda plataforma acontece quando os mainframes começam a perder força no mercado corporativo e substituídos pelos computadores pessoais, rodando aplicações no modelo cliente-servidor e conectadas à internet. No modelo cliente-servidor temos uma aplicação (cliente) rodando em um sistema final fazendo requisições de serviços ou recursos para um outro programa funcionando em outro sistema final (servidor) responsável por entregar o que está sendo solicitado (Kurose & Ross, 2010).

A era dos mainframes iniciada no pós-guerra faz parte da Primeira Plataforma Computacional, seguida pela Segunda Plataforma, baseada nos computadores pessoais, modelo cliente-servidor e na Internet. Estamos na terceira grande revolução dessa indústria, na qual dispositivos móveis, conexão constante com a Internet, interações sociais por meios digitais, grande capacidade de armazenamento e processamento a baixo custo são fatores destaque. Segundo a IDC (2014), a Terceira Plataforma Computacional é baseada em quatro principais pilares: Computação em Nuvem, Web Social, Big Data e Mobilidade. O objetivo deste trabalho é discutir os pilares tecnológicos da Terceira Plataforma Computacional, bem como os impactos sociais e possibilidades de novos empreendimentos a partir dessa evolução tecnológica.

Este documento segue assim estruturado: A Seção 2 introduz as tecnologias basais da Terceira Plataforma. A Seção 3 discute brevemente a mudança nas empresas a partir dos avanços tecnológicos. A Seção 4 apresenta uma análise de uma empresa inserida no contexto tecnológico atual e, por fim, na Seção 5 constam algumas conclusões e perspectivas da pesquisa.

2. Pilares Tecnológicos da Terceira Plataforma

A atual revolução tecnológica é baseada em quatro principais pilares: Computação em Nuvem, Web Social, Big Data e Mobilidade, brevemente aqui discutidas:

Computação em Nuvem: uma nova maneira de pensar e utilizar recursos computacionais, na qual o conteúdo que antes ficava no servidor físico da empresa, agora está em um alto nível de abstração, está na nuvem. A Nuvem consiste, segundo (Armbrust et al, 2010), em grandes conjuntos de servidores e demais dispositivos formando uma base responsável por manter os dados e serviços que serão oferecidos por meio da Internet.

Big Data: Conjunto de dados cujo tamanho está além da capacidade de ferramentas de bancos de dados tradicionais de capturar, armazenar, gerenciar e analisar (Manyika et al, 2011). Apesar dessa definição tratar principalmente sobre o tamanho do volume dos dados, big data está muito mais relacionado às características dos dados do que ao seu tamanho em si. Geralmente o processamento de big data está relacionado a duas necessidades das empresas: processamento para gestão dos dados e processamento para análise (analytics) (Géczy, 2014).

Web Social: Interações sociais não se dão mais somente no mundo real, mas também pela internet. As ferramentas sociais são responsáveis por gerar toneladas de dados diariamente, que podem se transformar em informações importantes para empresas acerca de seus consumidores. Tantos as redes sociais “públicas” (Boyd e

Ellison, 2007) quanto aquelas corporativas (Leonardi, Huysman e Steinfield, 2013) são partes fundamentais e uma das bases da terceira plataforma e formam o que chamamos de social business ou negócios sociais. Ambas as facetas do social business estão fortemente ligadas aos outros pilares da terceira plataforma. As redes sociais e seu turbilhão de dados são uma das principais responsáveis pela explosão de big data nos últimos tempos, o que também está ligado a popularização dos aparelhos móveis com acesso à internet.

Mobilidade: O ato de estar conectado a todo tempo e em qualquer lugar juntamente com a Nuvem permite que ferramentas sociais empresariais sejam usadas fora dos ambientes físicos das empresas e junto a seus funcionários o tempo todo. Livingston (2014) define computação móvel como uma tecnologia que nos permite transmitir dados, áudios, vídeos por meio de um computador ou qualquer dispositivo sem estar conectado a um local físico.

O conceito chave da Terceira Plataforma Computacional é o uso integrado das quatro tecnologias que definem seus pilares criando soluções de alto valor para a indústria. Gartner (2012) se refere à Terceira Plataforma como "The Nexus of Four Forces", descrevendo como a convergência e atração de Web Social, Mobilidade, Computação em Nuvem e Informações (Big data) juntas estão criando novas oportunidades de negócios e mudando o comportamento dos usuários. A Tabela 1 traz uma síntese de diferenças entre a segunda e terceira plataformas.

Tabela 1. Variables to be considered on the evaluation of interaction techniques

	Segunda Plataforma	Terceira Plataforma
Arquitetura	Tecnologia física local, rede Cliente-Servidor, Sistemas parrudos, completos e com mais possibilidade de customização.	Grandes Hardwares como responsabilidade do fornecedor, comunicação através da Internet e Sistemas variando de acordo com o contratado: IaaS, PaaS e SaaS.
Hardware	Grandes e caros servidores, menos acessíveis	Clusters, commodities, mais acessíveis
Segurança e Privacidade	Rede Fechada, TI interna responsável	Comunicação através da internet, protocolos
Pequenas e Médias Empresas	Alto custo, não acessível	Serviço sobre demanda, acessível
Dados	Estruturados	Não estruturados

Os aspectos de segurança e privacidade são também importantes para as organizações pesarem ao planejarem seus sistemas e produtos com a computação em nuvem. Na modelo *on-premise* largamente utilizado na era da Segunda Plataforma, os riscos de segurança eram menores (Claybrook, 2012), pois as empresas possuíam total controle sobre o sistema, podendo tratar suas necessidades de segurança e privacidade internamente. Já a oferta de soluções de aplicações através da internet no modelo de computação em nuvem traz um menor controle para as empresas dos aspectos de segurança e também traz novos riscos para esse ambiente, como a preocupação com o controle dos dados. Além disso, também existem as preocupações com a comunicação e

troca de informações entre o ambiente na nuvem e as empresas, visto que ela se dá toda através da internet, o que traz à tona mais riscos.

3. As Empresas e a influência da Terceira Plataforma

A evolução da tecnologia obriga a maioria dos negócios a se transformarem. Durante a primeira e segunda plataforma, as empresas foram moldadas a partir de grandes mainframes e servidores, onde a organização deveria fazer toda manipulação de seus recursos. Bancos são excelentes exemplos de sistemas que nasceram na Primeira e Segunda Plataforma e têm a necessidade de se adaptar a essa nova era para oferecer melhores serviços e experiência de uso a seus clientes. Mainframes ainda são figuras presentes nas arquiteturas computacionais dessas instituições, pela alta confiabilidade e complexo processo de migração de tecnologias. Entre as suas iniciativas estão acessar seus dados bancários, controlar cartões, investimentos e várias outras funcionalidades de sua conta pela Internet. Além da opção de acesso por desktop, os bancos têm também permitido acesso via dispositivos móveis. Empresas de varejo são também exemplos de adaptação, que percebendo o grande uso de aplicativos e Internet para efetuar compras, passou a desenvolver aplicativos para realizar vendas *online*, dar descontos, fazer promoções e ofertas direcionadas à necessidade do cliente a partir de seu histórico de navegação.

As empresas nascidas nos últimos 10 anos estão fazendo uso pleno de todo o potencial da Terceira Plataforma para alavancar seus negócios. Utilização da nuvem para uma melhor escalabilidade de seus produtos a menores custos, prover boa experiência do usuário em aparelhos móveis via aplicativos, fazer uso da colaboração e relação entre usuários ou coletar e processar dados por meio de um número enorme fontes são exemplos de como essas empresas estão inseridas nesse novo paradigma. Grande parte das empresas que consideramos altamente inovadoras na atualidade tem seus produtos e negócios fortemente baseados nos quatro pilares que compõem a Terceira Plataforma, a exemplo de Uber, Airbnb e Facebook. O que essas empresas têm em comum? Todas já nasceram na era da Terceira Plataforma e têm seus negócios fortemente fundamentos em computação em nuvem, mobilidade, *web* social e *big data*.

O Uber é uma organização que tem como produto final a locomoção de pessoas e é similar aos táxis tradicionais, porém a empresa não possui nenhum veículo, apenas faz o intermédio do serviço por um aplicativo. Com essa estratégia, o Uber alcançou um patamar maior do que qualquer outra empresa de táxi do mundo.

O Facebook também é um bom exemplo de negócio que surgiu na era da Terceira Plataforma e deu certo. Com o início da popularização dos celulares, o Facebook se inseriu nas plataformas móveis tornando disponível seu serviço em qualquer lugar com rede. A empresa construiu sua base em cima dos pilares: é uma rede social em que sua essência é integrar pessoas e compartilhar informações; está disponível através de computação em nuvem; a mobilidade está inserida quando temos aplicativos bem desenvolvidos para qualquer dispositivo móvel (Ha, 2015); e por último, a geração massiva de dados e informações.

Nesta nova era existe um novo modelo de negócios onde recursos são oferecidos como serviços, fazendo com que os clientes paguem sobre demanda, ou seja, paguem somente aquilo que utilizarem (Rouse, 2005). Neste modelo de negócio o preço será proporcional ao tamanho da empresa e sua utilização. Sendo assim, tecnologias que

antes eram muito caras, agora são acessíveis, possibilitando que empresas menores possam ser mais competitivas (Mohamed, 2009).

4. Spotify em Foco

O objetivo dessa seção é mostrar como a empresa Spotify está completamente inserida na Terceira Plataforma Computacional. Essa análise foi desenvolvida tomando como base artigos disponíveis na internet sobre a empresa e seus aspectos tecnológicos, entrevistas de funcionários da empresa, notícias relacionadas e apresentações feitas por engenheiros da empresa em eventos.

Spotify é uma empresa sueca que provê um serviço de *streaming* de músicas fundada em 2006 por Daniel Ek e Martin Lorentzon e que teve o lançamento oficial de seu produto em 2008. Conta com mais de 20 milhões de assinantes e mais de 75 milhões de usuários ativos (The Spotify Team, 2015). Em setembro de 2009 foi lançada a primeira versão do Spotify para dispositivos móveis, disponibilizada na Apple Store (iOS) e na Play Store (Android), somente para usuários Premium. Além de conseguir executar as faixas e *playlists* em qualquer lugar, os usuários estão aptos a usar o serviço de modo *offline*, baixando o conteúdo e utilizando mesmo desconectados da internet.

A Spotify surgiu numa época de crise da indústria da música onde o grande *player* do mercado de música digital era o iTunes, da Apple. Através de estratégias como permitir aos usuários ouvir as músicas sem pagar de maneira prática, tornar fácil a descoberta e compartilhamento de músicas e a dar a possibilidade de o usuário ter suas músicas locais armazenadas no serviço (Gopinath, 2014) fez com que a Spotify tivesse um crescimento meteórico, ultrapassando seus concorrentes e revolucionando a indústria da música.

Desde quando entrou em produção em 2008, Spotify tem como essência do seu produto a entrega de música através de streaming utilizando a internet, ou seja, entregar música sobre demanda. Esse é um modelo onde a empresa torna disponível todas as músicas a todo tempo, entregando pelo seu aplicativo o conteúdo que foi requisitado pelo usuário, que no caso é a música que ele quer ouvir (MaaS - *Music as a Service*), (Doerr *et al*., 2010). Sua estrutura era formada por servidores próprios *on-premise* e pela nuvem da Amazon, onde basicamente a nuvem armazenava os arquivos de áudio e os servidores locais, o restante dos dados incluindo o "*core*" da empresa (Konrad, 2016). A fim de obter benefícios da nuvem, como escalabilidade, o Spotify fez uso do Amazon S3 (*Amazon Simple Storage Service*) para armazenar os arquivos de áudio que eram acessados pelos usuários.

Através desses servidores, a empresa consegue prover as funcionalidades de pesquisa de músicas, listas de reprodução, funções sociais e execução de faixas musicais. Quando um usuário ouve uma música, esses dados são entregues por meio de uma melhor combinação: se a música foi ouvida recentemente, o Spotify a executará a partir da cache; através do modelo P2P (*peer-to-peer*), acessando os dados providos por outros usuários que ouviram essa música; se nenhuma das duas iniciais for possível, o usuário obterá a faixa provida pelo servidor. Com esse modelo, o acesso ao servidor é reduzido, e a forma de entrega é otimizada (Yanggratoke *et al.*, 2013).

Spotify utilizou P2P compartilhando dados entre os usuários, mas segundo (Sar, 2014), a prática não era válida para dispositivos móveis, e somando-se ao grande número de servidores que já possuía, em 2014 deixaram de usar P2P, centralizando toda

recuperação de informação em servidores próprios. De acordo com (Harteau, 2016), em fevereiro de 2016 foi anunciada a mudança da base de infraestrutura do Spotify para a nuvem do Google, a *Google Cloud Plataform*, passando, segundo (Leygues, 2016), a usar também ferramentas como *Google BigQuery* e *Google Cloud DataFlow*, ganhando também em análise de dados para tomada de decisão.

O Spotify é uma empresa *data driven*, ou seja, uma empresa que faz uso dos dados em praticamente todas as suas áreas para a tomada de decisões sempre que possível (Spotfire Blogging Team, 2015). O grande volume de dados produzido pelos seus usuários abre um leque de oportunidades para empresa gerar valor através de sua análise. Esse grande volume de dados é usado para os mais diferentes fins dentro na empresa. Abaixo vamos dar alguns exemplos de como essa cultura, voltada para dados, influencia nas diferentes áreas da empresa:

- **Pagamento de direitos**: Spotify paga aos detentores dos direitos da música por vez que a música é reproduzida no serviço. Portanto, precisam contabilizar sempre que cada uma das músicas presentes no serviço for tocada, o que afeta diretamente o setor financeiro da empresa, que exige precisão dos dados gerados.
- **Gerar mais engajamento**: é importante que os assinantes usem o serviço e vejam sempre valor agregado para que não cancelem sua assinatura, portanto, precisam de iniciativas para gerar engajamento do consumidor e precisam conseguir analisar seu impacto por meio de métricas.
- **Recomendar melhor**: através da captura e análise dos dados de consumo musical dos seus usuários, Spotify consegue fazer recomendações de músicas que tenham a ver com o que cada usuário gosta. A "Descoberta da semana", uma *playlist* criada para cada usuário do serviço com músicas personalizadas para seu gosto, em apenas 10 semanas atingiu 1 bilhão de faixas tocadas e fez com que 71% dos usuários adicionassem pelo menos umas das músicas ali presentes nas suas *playlists* pessoais (Spotify, 2015). Fazer essas recomendações com eficácia só é possível pela análise de *big data*.

Spotify é uma empresa que dificilmente surgiria na era da Segunda Plataforma, pois o uso de dados (no caso, de um grande volume deles) é uma base para as tomadas de decisões e só com o desenvolvimento de técnicas e ferramentas de *big data* é possível realizar tudo que o Spotify faz. Podemos afirmar que o uso de *big data* é um dos grandes pilares da companhia que ela faz pleno uso desse pilar da Terceira Plataforma no seu dia-a-dia.

Além da praticidade de ter um catálogo enorme de músicas de maneira fácil, outra característica está nas possibilidades de interação social na plataforma. É possível seguir seus amigos/*playlists*/artistas, ver o que estão ouvindo, criar *playlists* compartilhadas entre outras funções que proveem uma experiência social a um serviço que tem como principal função permitir que as pessoas ouçam músicas.

Todo o sistema de interações sociais do Spotify faz uso do paradigma de *Publish/Subscribe* (publicar/assinar em tradução livre). Esse modelo é baseado em eventos, onde "assinantes" podem expressar seu interesse em um evento ou padrão de eventos e são posteriormente notificados caso algum evento, gerado por um "publicador" se encaixe naquele interesse manifestado. Nesse paradigma, Spotify utiliza em seu serviço o modelo pub/sub baseado em tópicos, ou seja, o usuário pode se

inscrever ou seguir tópicos que, conforme (Setty *et al.*, 2013) são os seguintes: amigos, *playlists* e páginas de artistas.

As funções sociais presentes no Spotify se encaixam em todos os quesitos da definição que Boyd e Ellison (2007) dão para redes sociais. Mesmo com o seu negócio fim não sendo uma rede social, faz forte uso das características de uma para entregar ao usuário maiores possibilidades de interação e ter outras funções no seu serviço. Os aspectos sociais são parte fundamental do serviço e um pilar da Terceira Plataforma muito bem aproveitado para agregar valor ao produto oferecido.

Tabela 2. Spotify e a Terceira Plataforma

	Terceira Plataforma	Aspectos	Spotify
Computação em Nuvem	Hardware hospedado pelo fornecedor; Produto oferecido como um serviço; escalável;	Infraestrutura	Hospedado na nuvem pública do Google;
		Escalabilidade	Possibilidade de expandir recursos mediante uma grande demanda
Big Data	Capturar, armazenar, gerenciar e analisar um grande volume de dados verossímeis, de grande variedade e com grande velocidade;	Volume	Gera mais de 14 TB dados de log/dia;
		Processamento e Análise	Captura, processamento e análise dos dados de uso dos assinantes para funções como pagamento de direitos e recomendações de músicas
Web Social	Serviço Web, que permite construir um perfil, articular, ver e examinar/cruzar conexões;	Articular lista de Conexões	Possível seguir amigos/artistas;
		Postar/Compartilhar	Ver suas atividades e de amigos, criar playlists compartilhadas;
Mobilidade	Tecnologia que permite transmitir dados, áudios, vídeos por meio qualquer dispositivo sem estar conectado a um local físico;	Alcance	Disponível através de aplicativo para Android e iOS, para contas pagas e não pagas; +50% dos acessos através de dispositivos móveis

A Tabela 2 resume os aspectos da Terceira Plataforma identificados no Spotify, uma empresa inteiramente inserida neste paradigma. Em seu início a empresa possuía uma solução híbrida, onde utilizava recursos *on-premise* com computação em nuvem, sem inclusão de dispositivos móveis e usando ferramentas tradicionais como Hadoop, MapReduce e Hive. É nítido que o sistema evoluiu, se aproximando cada vez mais das últimas tecnologias lançadas no mercado. Fica evidente a utilização e busca de cada vez mais explorar a computação em nuvem, interações sociais e a mobilidade.

5. Conclusão

Estamos em uma fase de transição, onde o hoje forma a base da tecnologia de um futuro próximo e muito diferente, habilitando novos paradigmas, criando mercados, novas formas de fazer negócio, novos serviços e empregos.

Com esse trabalho temos o objetivo de contribuir com um melhor entendimento do que é a Terceira Plataforma Computacional e de como ela pode trazer grandes oportunidades para as empresas. Foram expostos os aspectos relevantes que devem estar presentes na Terceira Plataforma e essa exploração pode ser continuada de várias formas, como: detalhando a especificação de requisitos, pesquisando as melhores práticas para desenvolvimento, melhores práticas para migração da Segunda para

Terceira Plataforma, levantando e analisando os aceleradores da Terceira Plataforma (Internet das Coisas, Robótica, Sistemas Cognitivos, etc).

Referências

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., et al. (April de 2010). A View of Cloud Computing. Communications of the ACM, 53.

Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship.

Claybrook, B. (2012, Julho). On-premises vs. SaaS: Making the choice. Retrieved from http://searchcloudapplications.techtarget.com/tutorial/On-premises-vs-SaaS-Making-the-choice

Doerr, J., Benlian, A., Vetter, J., & Hess, T. (2010). Pricing of Content Services – An Empirical Investigation of Music as a Service.

Gartner (2012). The Nexus of Forces: Social, Mobile, Cloud and Information.

Géczy, P. (2014). Big Data Characteristics. The Macrotheme Review.

Gopinath, S. (2014). The Third Disruptor: Spotify and the Future of Digital Music.

Ha, A. (2015). More Than Half A Billion People Access Facebook Solely From Mobile. Fonte: TechCrunch: https://techcrunch.com/2015/01/28/facebook-mobile-only-2/

Harteau, N. (2016). Announcing Spotify Infrastructure's Googley Future. Spotify News

IDC (2014). International Data Corporation: Predictions 2015: Accelerating Innovation — and Growth — on the 3rd Platform. International Data Corporation (IDC).

Konrad, A. (2016). Why Spotify Really Decided To Move Its Core Infrastructure To Google Cloud. Fonte: Forbes.

Kurose, J. F., & Ross, K. W. (2010). Redes de computadores e a Internet: uma abordagem top-down.

Leonardi, P. M., Huysman, M., & Steinfield, C. (2013). Enterprise Social Media: Definition, History, and Prospects for the Study of Social Technologies in Organizations. Journal of Computer-Mediated Communication.

Leygues, G. (2016). Spotify chooses Google Cloud Platform to power data infrastructure. https://cloudplatform.googleblog.com/2016/02/Spotify-chooses-Google-Cloud-Platform-to-power-data-infrastructure.html

Livingston, D. (2014). Introduction & History of Mobile Computing.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). Big data: The next frontier for innovation, competition, and productivity.

Mohamed, A. (Junho de 2009). The benefits of low-cost SaaS for SMEs. http://www.computerweekly.com/feature/The-benefits-of-low-cost-SaaS-for-SMEs

Rouse, M. (Setembro de 2005). What is metered services (pay-per-use)? - TechTarget: http://searchcio.techtarget.com/definition/metered-services

Sar, E. (2014). Spotify Starts Shutting Down its Massive P2P Network torrentfreak.com/spotify-starts-shutting-down-its-massive-p2p-network-140416/

Setty, V., Kreitz, G., Vitenberg, R., Steen, M. v., Urdaneta, G., & Gimåker, S. (2013). The Hidden Pub/Sub of Spotify (Industry Article).

Spotfire Blogging Team. (2015). What it Means to Be a Data-Driven Enterprise. http://www.tibco.com/blog/2015/03/31/what-it-means-to-be-a-data-driven-enterprise/

Spotify. (2015). Discover Weekly Reaches One Billion Tracks Streamed in 10 Weeks. https://news.spotify.com/us/2015/10/08/discover-weekly-reaches-one-billion-tracks-streamed-in-10-weeks-2/

Yanggratoke, R., Kreitz, G., Goldmann, M., Stadler, R., & Fodor, V. (2013). On the performance of the Spotify backend. Network and Systems Management.

Representação das correntes do trabalho escravo através de Linked Open Data

Leticia Verona

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

leticiaverona@ufrj.br

Abstract. *The object of this study is the modeling and representation process of the Dirty List of Slavery Records in Brazil in Linked Open Data formats. The list is published by the Ministery of Labor since 2004. This work aims to make the list available in high standard form, respecting all the principles of W3C Open Linked Data.*

Resumo. *O objeto deste estudo é o processo de modelagem e representação da Lista Suja do Trabalho Escravo no Brasil em formato de Dados Abertos Ligados. A lista é publicada pelo Ministério do Trabalho desde 2004. Este trabalho objetiva tornar a lista disponível usando padrões e respeitando todos os princípios dos dados abertos ligados do W3C.*

1. Introdução

O objetivo deste trabalho é trazer à luz as conexões entre aqueles flagrados utilizando trabalho escravo e o poder público no Brasil. A hipótese, confirmada pelos dados, é que o trabalho escravo sobrevive no Brasil porque os seus agentes possuem conexões próximas com o poder e, através destas conexões, garantem a impunidade e a continuidade dos seus atos. Por outro lado, o poder econômico, calcado no sofrimento destes escravos contemporâneos, financia e perpetua a participação destes atores na política. Segundo [Figueira 2014]: "Se a omissão das autoridades torna mais grave a história dessa gente, ela pode ainda ser pior, se a autoridade for proprietária de imóvel denunciado". Esta conclusão, bastante óbvia, não seria um fruto significativo por si só, pois o leitor atento e dedicado poderia obter as informações sozinho. Além disso, análises sobre a cadeia produtiva ligada ao trabalho escravo no Brasil podem ser encontradas em repositórios de entidades de combate ao trabalho escravo, como a Pastoral da Terra, a ONG Repórter Brasil e o grupo de pesquisa sobre Trabalho Escravo Contemporâneo (GPTEC-UFRJ), entre outros.

Dessa forma, o objeto resultante deste estudo não são as conexões encontradas, mas sim sua modelagem e publicação segundo padrões de *Linked Open Data*, tornando as informações sobre os senhores de escravos brasileiros parte da Web Semântica e permitindo que o potencial desta informação seja multiplicado. O conjunto de dados gerado não é grande mas, neste caso, é preciso concordar com [Boyd and Crawford 2012] quando dizem que devemos reconhecer a importância de "pequenos dados" e que intuições para pesquisa e compreensão da realidade podem ser encontradas em escalas muito modestas. Os mesmos autores ressaltam também que independente do tamanho do dado, ele está sujeito a limitações e vieses. Para evitar a má interpretação, especialmente no caso de dados

tão impactantes, este estudo busca explicitar ao máximo o contexto de cada informação através de uma modelagem estrita e embasada.

Um trabalho similar foi realizado no Chile, com a criação do site *Poderopedia* [Hernández 2013] que usa os dados abertos ligados para exibir as relações de poder entre políticos e empresários. No Uruguai, os dados do governo foram utilizados para tornar disponível ao público o tempo de espera e qualidade de atendimento de cada uma das unidades públicas de saúde [Köster and Suárez 2016].

A necessidade de dar significado semântico e publicar os dados governamentais foi discutida em [Tygel et al. 2016], que propõe a utilização de uma camada intermediária para conciliar *tags* utilizadas, unificando vocabulário e promovendo a disseminação dos dados. Os autores afirmam que a confusão de formatos e nomenclaturas e principalmente a ausência de metadados têm como resultado que a navegação, exploração e busca dentro dos portais de dados abertos e especialmente na interligação entre eles é dificultada, senão impedida. O restante do artigo está organizado da seguinte forma. Na seção 2, são definidos os conceitos do trabalho escravo contemporâneo e como é formada a Lista Suja do Trabalho Escravo no Brasil, o nó base da rede de informações gerada por este estudo. Na seção 3, são apresentados alguns conceitos fundamentais sobre dados governamentais abertos. Na seção 4, são explicados o método de obtenção da informação, as decisões de modelagem e um modelo para visualização dos dados. Por fim, na seção 5, são apresentadas as conclusões e sugestões de trabalhos futuros.

2. Trabalho Escravo Contemporâneo no Brasil

Segundo a Comissão Pastoral da Terra (CPT), "Trabalho Escravo contemporâneo é a sujeição física ou psicológica de um homem por outro. No caso brasileiro o instrumento mais comum de sujeição é a dívida crescente e impagável". Frequentemente, os trabalhadores sujeitos ao trabalho escravo são pessoas em situação vulnerável, a quem o aliciador oferece promessas de uma vida melhor, muitos vezes reforçada por um adiantamento financeiro. O mecanismo é descrito por [Figueira 2005]:

> *"Com esse objetivo é construído um sistema de endividamento progressivo do trabalhador. A dívida começa quando, ao ser contatado, o peão recebe do gato ou de um seu preposto um pequeno adiantamento em dinheiro. E aumenta a dívida com os gastos de transporte e alimentação até a unidade de produção. Mas o ciclo de endividamento não termina aí. Ele prossegue nas compras de alimentação, material de higiene, ferramenta de trabalho, instrumento de proteção e medicamento feitas na cantina do empreiteiro ou da empresa proprietária da fazenda. Desinformado de seus direitos, o trabalhador tem uma consciência falsa de responsabilidade legal e moral sobre a 'dívida'. Impulsionado pela noção de que 'quem deve é obrigado a pagar'; torna-se primeiro prisioneiro de sua própria consciência, pois desconhece que no Brasil ninguém é obrigado a trabalhar ou é preso por dívida, salvo nos casos específicos de omissão paterna ou materna em pensão alimentar. Depois se torna prisioneiro da distância, da falta de dinheiro para tomar um transporte, da vergonha de retornar à casa mais pobre do que saiu, ou pelas ameaças e por homens armados."*

O combate ao trabalho escravo no Brasil ainda carece de avanços significativos, apesar de esforços notáveis terem sido feitos nos governos Fernando Henrique, Lula e Dilma. Uma das mais poderosas ferramentas, criada pela portaria de n. 504 do Ministério do Trabalho e Emprego em 2004, é a Lista Suja do Trabalho Escravo. Como complemento a ela foi criada a portaria n. 1.150, do Ministério da Integração Nacional no mesmo ano. Como resumido por [Viana 2007], a primeira criou o cadastro de pessoas físicas e jurídicas que exploram o trabalho "em condições análogas a de escravo". A segunda recomenda aos órgãos financeiros que não lhes concedam regalias. O autor ainda ressalta que a inserção na lista depende de não caber mais recurso administrativo, no qual se assegura ampla defesa.

Desta forma, dentro dos mais nobres preceitos constitucionais, a lista, ao mesmo tempo que publica as informações, retira dos agentes do trabalho escravo uma de suas maiores fontes de riqueza, a obtenção de crédito rural nos bancos públicos. O governo combate e deixa de financiar o trabalho escravo, pelo menos formalmente.

A Lista Suja enfrenta muitos opositores, dentro e fora do governo, e frequentemente é retirada dos portais devido a liminares e mudanças nas regras. Para efeito deste estudo, utilizamos a versão publicada pela ONG Repórter Brasil [Brasil 2016], publicada em XML e com os seguintes metadados explicitados:

- **Proprietário:** nome do empregador;
- **CPF-CNPJ-CEI:** CPF, CNPJ ou CEI do empregador, somente números;
- **Estabelecimento:** nome do estabelecimento;
- **Endereço:** endereço do empregador;
- **Cidade:** cidade do empregador;
- **Estado:** estado do empregador;
- **Libertados:** número de trabalhadores libertados na operação;
- **Inclusão:** data da inclusão na Lista Suja.

3. Linked Open Data e dados governamentais abertos

Em 2009, o Brasil participou da fundação do movimento *Open Government* e, em conjunto com outros sete países, assinou o primeiro tratado de transparência internacional, encabeçado pelos Estados Unidos e posteriormente endossado por mais de 50 países. Estes países comprometeram-se a publicar seus dados para garantir o acesso livre à informação [Obama 2009]. Além de acessível, o dado aberto significa diminuir ao máximo as barreiras para o seu uso e reuso. Segundo o tratado *Open Government*, o dado para ser considerado aberto deve ser compatível com os seguintes princípios: (i) ser completo, não sujeito a sanções de privacidade e limitações de privilégios; (ii) ser primário, no sentido de ter a menor granularidade possível; (iii) ser acessível; (iv) ser processável por máquinas; (v) ser sensível ao tempo, ou seja, publicado em tempo de ser útil; (vi) ter acesso livre, independente de qualquer registro ou aprovação; (vii) usar dados não proprietários e com licença livre.

Desde então, o portal de dados abertos do governo brasileiro (`dados.gov.br`) traz uma coleção bastante extensa de dados. Apesar de valiosos, os conjuntos de dados, em sua maioria, necessitam de tratamento prévio para serem utilizados com eficiência e permitirem a obtenção de conclusões analíticas a partir deles. Como expressamente

endereçado por [Bauer 2011], integrar informações de diferentes fontes é custoso em termos de tempo e dinheiro e a ideia básica da Web Semântica é criar maneiras eficientes de publicar a informação em ambientes distribuídos, reduzindo estes custos através do uso de padrões largamente disseminados.

O mapeamento de informações entre o receptor e o transmissor se dá através da sintaxe, esquemas e vocabulários usados para publicar o dados. Além de tudo, é fundamental fornecer informações de qualidade aos desenvolvedores e pesquisadores sobre os dados, explicando sua proveniência e viabilizando que trabalhos de qualidade sejam feitos sobre eles.

O movimento "semântico" da Internet foi antecipado por [Kent 1978] em *Data and Reality*: "A necessidade de uma abordagem mais sofisticada de descrição dos dados vai crescer na medida que as interfaces dos sistemas se expandirem e envolverem pessoas que não são treinadas em disciplinas computacionais". Outros visionários foram [McLuhan 1969] que anteciparam o problema da sobrecarga de informação: "As informações despencam sobre nós, instantaneamente e continuamente. Tão pronto se adquire um novo conhecimento, este é rapidamente substituído por informação mais recente. Nosso mundo, eletricamente configurado, forçou-nos a abandonar o hábito de dados classificados para usar o sistema de identificação de padrões.".

4. Método e Modelagem

O esforço de modelagem necessário para a publicação de dados nos moldes de *Linked Open Data* deve ser visto como uma etapa no caminho de desobstruir o acesso ao dado, uma vez que, modelando o dado para funcionar independente do contexto de uma aplicação específica, estamos preparando o mesmo para qualquer uso. A "receita" do *W3C Linked Data Cookbook* [Hyland and Terrazas 2011] foi o guia adotado para a modelagem da Lista Suja do Trabalho Escravo como *Linked Data*. Este guia apresenta as melhores práticas para produção de *Linked Data* desde a sua modelagem até a publicação dos dados em si. Na etapa de modelagem, os autores resumem os seguintes passos principais: identificar, modelar, nomear e testar. As subseções a seguir explicitam cada um desses passos instanciados para a modelagem da Lista Suja.

4.1. Identificar: Identificação das entidades e das ligações entre elas

Neste passo, devem ser buscados, nos dados, quais são as entidades de interesse. No caso deste estudo, foram identificadas as seguintes entidades primárias na Lista Suja e ligações entre elas:

- **Lista Suja do Trabalho Escravo:** lista de operações de flagrante de trabalho escravo;
- **Operação:** operação que levou o proprietário a estar na lista. Possui os atributos *nome do estabelecimento*, *endereço do estabelecimento*, *número de empregados libertados* e *data de inclusão na lista*;
- **Cidade:** cidade onde se efetivou a operação;
- **Proprietário:** pessoa ou organização dona do imóvel onde foi flagrado o trabalho escravo. Esta entidade possui o atributo *documento*.

Nesta identificação de entidades duas decisões de modelagem podem ser ressaltadas: (i) alguns dados foram considerados apenas atributos de entidades e (ii) a informação

do estado foi suprimida, pois foi considerada uma informação secundária, obtida a partir da cidade.

4.2. Modelar e Nomear: Seleção de vocabulários e ligações com dados já existentes

A modelagem envolve primeiramente a decisão de que vocabulário deve ser usado para descrever cada entidade. Segundo as recomendações do W3C, deve ser selecionado, sempre que possível, um vocabulário amplamente disseminado.

Os vocabulários selecionados para modelagem da Lista Suja foram:

- **classe *list* da ontologia *RDF Schema* [Brickley and Guha 2004]:** selecionada para modelar a entidade *Lista Suja do Trabalho Escravo*;
- **classe *event* da *The Event Ontology*[Raimond and Abdallah 2007]:** selecionada para modelar a entidade *Operação*;
- **classe *adm2* da ontologia *Geonames*[Vatant and Wick 2012]:** selecionada para modelar a entidade *Cidade*;
- **classe *Person* da ontologia *FOAF* [Brickley and Miller 2012] e classe *Organization* da ontologia *The Organization Ontology* [Dave Reynolds 2014]:** selecionadas para modelar a entidade *Proprietário* sendo a primeira para o caso do proprietário ser uma Pessoa Física e a segunda para Pessoa Jurídica.

Algumas ocorrências da entidade *Proprietário*, quando não são Pessoas Físicas, necessitaram de uma investigação mais aprofundada para enumeração dos seus responsáveis. Quando se trata de uma filial, é necessário identificar sua matriz, pois a partir da matriz donos e diretores podem ser enumerados. Neste caso, foi utilizada a classe **SubOrganizationOf** da ontologia *The Organization Ontology* [Dave Reynolds 2014].

No caso da entidade *Proprietário* ser uma Pessoa Jurídica, foram procurados os dados no site da Receita Federal (`receita.fazenda.gov.br`) para identificar os responsáveis pela organização. Desse modo, a relação entre a organização e a pessoa responsável deve ser reificada. Como diz [Guarino 2016], relações reificadas são fundamentais para endereçar muitos dos clássicos problemas de modelagem, onde é necessário descrever atributos do relacionamento como a sua extensão no tempo ou evolução de papéis. No caso dos responsáveis pelas empresas constantes da Lista Suja, a pessoa pode assumir vários papéis neste relacionamento que a implicam com o trabalho escravo, tais como: sócia da empresa (dado obtido na Receita Federal para empresas do modelo sociedade limitada), diretora (dado obtido na Receita Federal para empresas do modelo sociedade anônima), presidente (dado obtido na Receita Federal no caso de empresas do modelo federação).

Outro dado interessante, porém não disponível com facilidade, é se a pessoa é acionista de uma sociedade anônima. Foi utilizado o conceito de "papéis" para identificar como a pessoa se relaciona com a empresa através da classe **Role** da ontologia *Organization* [Dave Reynolds 2014].

Após a seleção dos vocabulários, é recomendado no *W3C Linked Data Cookbook* que seja descoberto se outros já descreveram dados similares e seja decidido se serão ou não feitos *links* com as entidades já descritas em outros conjuntos de dados. Aqui existem dois caminhos: (i) pode-se criar uma entidade própria e usar um predicado de ligação como o *owl:sameAs*, ou (ii) indicar diretamente a entidade já descrita (através do uso da URI existente).

No caso dos dados da Lista Suja do Trabalho Escravo, a entidade *Operação* foi ligada diretamente à entidade *Cidade* do conjunto de dados GeoNames (`geonames.org`). No caso da entidade *Proprietário*, a decisão foi pela criação de uma entidade própria e, quando a Pessoa ou Organização foi localizada no conjunto de dados da DBpedia (`dbpedia.org`), foi criada uma ligação utilizando o predicado *owl:sameAs*.

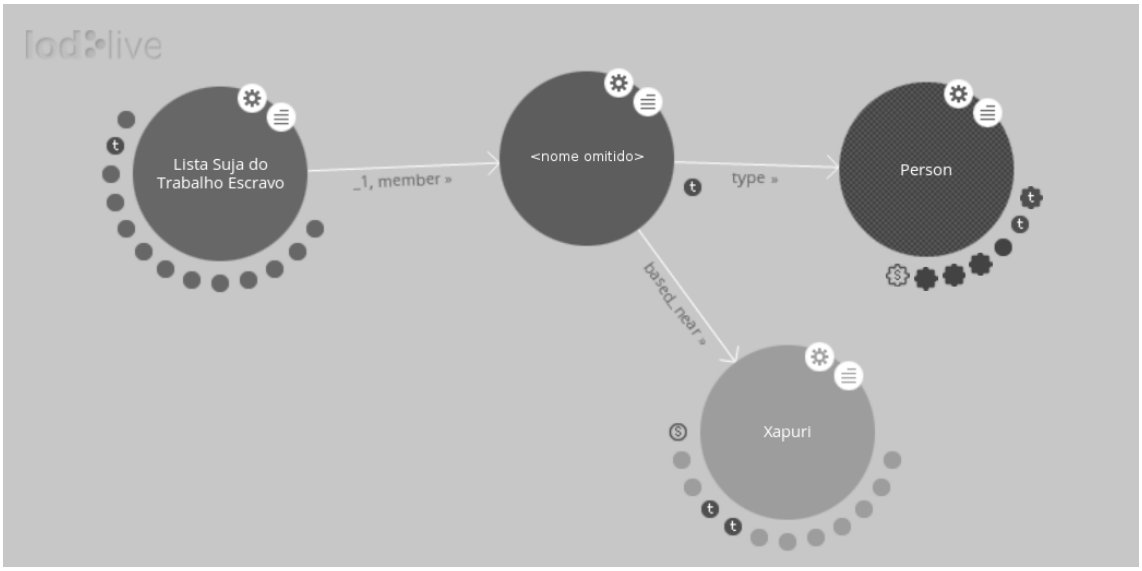


Figura 1. Exemplo da visualização de um nó com a entidade do tipo pessoa física modelada como *foaf:Person*.

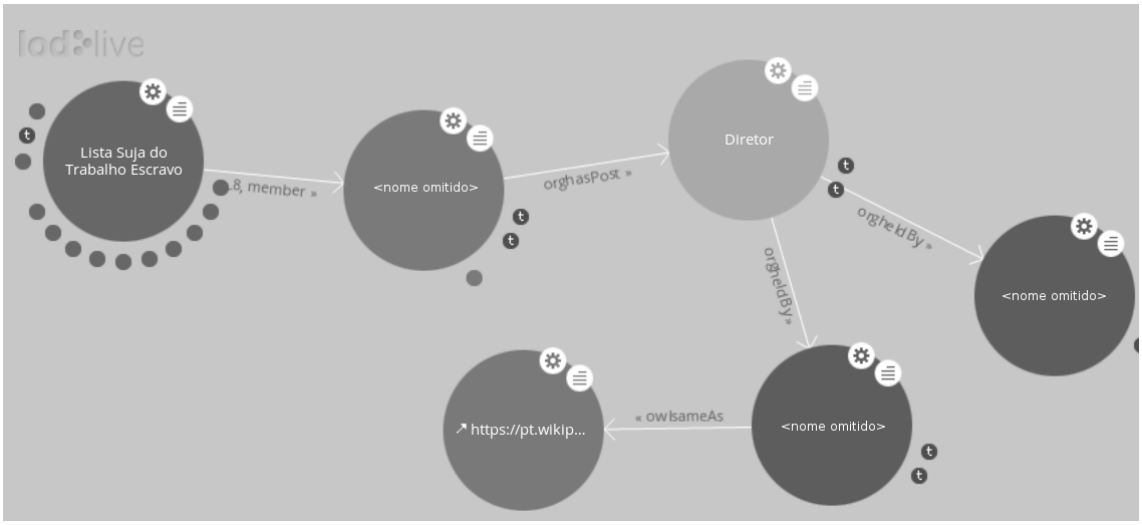


Figura 2. Exemplo da visualização de um nó com a entidade do tipo pessoa jurídica expandida até que a pessoa física seja identificada.

4.3. Testar: Visualização

Para testar as assunções feitas na modelagem é sugerido no *W3C Linked Data Cookbook* que seja feita a construção de diagramas dos objetos e relacionamentos para comunicar-se rapidamente quais itens e relacionamentos estão refletidos no modelo gerado. Além disso, [Bauer 2011] afirmam ser fundamental criar ferramentas poderosas sobre os dados,

que permitam ao usuário final, através de interfaces amigáveis, analisar e consumir os seus dados. Para facilitar a visualização da Lista Suja, a ferramenta LodLive (`lodlive.it`) foi utilizada para auxiliar na análise e exploração dos dados obtidos. Através da ferramenta, que viabiliza a navegação entre entidades através de *SPARQL endpoints*, foi possível descobrir, por exemplo, que proprietários de imóveis incluídos na Lista Suja são ligados ao poder público. Dois exemplos de visualização podem ser vistos nas figurass 1 e 2. Para efeito deste artigo, omitimos os nomes constantes na lista.

5. Conclusões e Trabalhos Futuros

O presente artigo descreveu como o uso de padrões para publicação de dados abertos ligados pode facilitar o acesso e a interligação de dados visando pontencializar o consumo desses através de diferentes aplicações e possibilitar análises integradas desses dados advindos de diferentes fontes. Um dos passos fundamentais para atender aos requisitos para dados abertos ligados é publicar o conjunto de dados em um domínio e anunciar esta publicação, considerando a responsabilidade social de um publicador de dados [Hyland and Terrazas 2011]. Um conjunto de dados ligados, após publicado, não pode ficar indisponível sem o risco de "afetar" aplicações de terceiros (que esteja consumindo seus dados).

Além disso, conforme destacado já na introdução deste artigo, o objeto resultante deste estudo é a modelagem e publicação da Lista Suja do Trabalho Escravo nos padrões de *Linked Open Data*, tornando as informações sobre os senhores de escravos brasileiros parte da Web Semântica e permitindo que o potencial desta informação seja multiplicado. Este artigo também apresentou exemplos de visualizações que podem ser obtidas através da exploração dos dados disponibilizados. Tais visualizações trazem à luz as conexões entre aqueles flagrados utilizando trabalho escravo e o poder público no Brasil, trazendo subsídios para confirmar a hipótese de que o trabalho escravo só sobrevive no Brasil porque os seus agentes possuem conexões próximas com o poder e, através destas conexões, acabam garantindo a impunidade e a continuidade desses atos.

Do ponto de vista da expansão do conjunto de dados modelado neste artigo, existem dados abertos disponíveis sobre doações de campanha, composição de câmaras legislativas e membros do poder executivo e judiciário que podem ser ligadas aos agentes que constam na Lista Suja do Trabalho Escravo. A inclusão de parentes diretos (pais, filhos, irmãos) e sócios poderia também trazer à luz muitos comprometimentos. Outra expansão seria o uso de ferramentas de descoberta de links, tais como LIMES [Ngomo 2011] e Silk [Bizer et al. 2009], para interligação dos dados presentes na Lista Suja aos de outros conjuntos de dados.

Este trabalho foi desenvolvido no contexto da disciplina de Organização do Conhecimento do Programa de Pós-graduação em Informática da UFRJ. Agradeço às profas. Giseli Lopes e Maria Luiza M. Campos pelas sugestões e revisões realizadas.

Referências

Bauer, Florian e Kaltenböck, M. (2011). Linked open data: The essentials. *Edition mono/monochrom, Vienna*.

Bizer, C., Volz, J., Kobilarov, G., and Gaedke, M. (2009). Silk - a link discovery framework for the web of data. In *18th International World Wide Web Conference*.

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679.

Brasil, R. (2016). Lista suja do trabalho escravo no brasil. disponível em: www.reporterbrasil.org.br/lista-suja/. *Acesso em: 22 de agosto de 2016*.

Brickley, D. and Guha, R. V. (2004). {RDF vocabulary description language 1.0: RDF schema}.

Brickley, D. and Miller, L. (2012). Foaf vocabulary specification 0.98. *Namespace document*, 9.

Dave Reynolds, E. L. (2014). The organization ontology w3c recomendation.

Figueira, Ricardo Rezende e Prado, A. A. (2014). Trabalhadores denunciam o trabalho escravo. *Hendu–Revista Latino-Americana de Direitos Humanos*, 4(1):22–40.

Figueira, R. R. (2005). A migração e o trabalho escravo por dívida no brasil. *Travessia na desordem global: Fórum Social das Migrações*, 1:181–189.

Guarino, Nicola e Guizzardi, G. (2016). Relationships and events: towards a general theory of reification and truthmaking. In *15th International Conference of the Italian Association for Artificial Intelligence (2016, submitted)*.

Hernández, D. R. (2013). Estándares de publicación de datos para la información pública en chile.

Hyland, B. and Terrazas, Boris Villazón e Capadisli, S. (2011). Cookbook for open government linked data. *W3C, W3C Task Force-Government Linked Data Group*.

Kent, W. (1978). *Data and reality: Basic assumptions in data processing reconsidered*. Elsevier Science Inc.

Köster, V. and Suárez, G. (2016). Open data for development: Experience of uruguay. In *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance*, pages 207–210. ACM.

McLuhan, Marshall e Fiore, Q. (1969). *O meio são as massa-gens*. Record.

Ngomo, Axel-Cyrille Ngonga e Auer, S. (2011). Limes-a time-efficient approach for large-scale link discovery on the web of data. *integration*, 15:3.

Obama, B. (2009). Transparency and open government. *Memorandum for the heads of executive departments and agencies*.

Raimond, Y. and Abdallah, S. (2007). The event ontology. Technical report, Citeseer.

Tygel, A., Debattista, J., Orlandi, F., Campos, M. L. M., et al. (2016). Towards cleaning-up open data portals: A metadata reconciliation approach. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pages 71–78. IEEE.

Vatant, B. and Wick, M. (2012). Geonames ontology.

Viana, M. T. (2007). Trabalho escravo e "lista suja": um modo original de se remover uma mancha. *Brasília: Organização Internacional do Trabalho*.

Aumento da Adesão e do Engajamento de Usuários do Campus Social com Uso de Mecanismos de Gamificação

Eliel Roger da Silva[1], Tiago Cruz de França[1,2], Jonice de Oliveira Sampaio[2]

[1]Departamento de Matemática – Universidade Federal Rural do Rio de Janeiro (UFRRJ)
Seropédica – RJ – Brasil
[2]Programa de Pós-Graduação em Informática – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

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

Abstract. *Many collaborative systems face challenges related to both gathering new users and keeping their users engaged using such systems. The same challenge is faced by the Campus Social the tool of our interest which support opportunistic communication. We have described our gamification approach as a proposal to leverage both the gain of new users and the engagement in using this application. Gamification has been rather adopted in mobile applications. Therefore, our proposal adapts mechanisms utilized by games to build a suitable approach for the Campus Social.*

Resumo. *Muitos sistemas colaborativos enfrentam dificuldades com a adesão de usuários e com seu engajamento com uso da ferramenta. A partir da observação dessa dificuldade enfrentada pela aplicação de comunicação oportunística utilizando dispositivos móveis chamada Campus Social, o presente trabalho apresenta uma abordagem de gamificação para aumentar a adesão e engajamento de usuários nessa aplicação. A gamificação tem sido uma técnica bastante adotada em aplicativos para dispositivos móveis. Por esse motivo a proposta apresentada adapta técnicas utilizadas em jogos eletrônicos ao contexto do Campus Social.*

1. Introdução

Observando as dificuldades cotidianas de comunicação e de falta de informação enfrentadas nos campi universitários brasileiros e enxergando o potencial oferecido pela comunicação sem fio e dispositivos móveis, surgiu o Campus Social [Tabak et al. 2015].

O Campus Social difere de outras mídias populares (Facebook[1] ou Google+[2], por exemplo) porque seu foco é atender especificamente o público universitário facilitando principalmente a comunicação dentro dos campi, tornando-a mais rápida, organizada e estimulando um maior diálogo entre diferentes linhas de raciocínio. Trata-se de uma abordagem para colaboração baseada em comunicação oportunística e dispositivos móveis. A comunicação oportunística é o modo de comunicação *ad hoc* definida, no contexto deste trabalho, como comunicação estabelecida eventualmente motivadas por um interesse comum.

Em outras palavras, o Campus Social é uma ferramenta que proporciona a criação de redes baseadas no interesse cotidiano de pessoas que frequentam o mesmo ambiente (como um campus universitário). Por exemplo, se um usuário tem uma dúvida sobre o refeitório ou biblioteca ele pode expor a sua dúvida a comunidade. De forma oportuna, algum outro usuário pode ter a informação e compartilhá-la de duas formas: respondendo diretamente ou apontando postagem prévia que sirva como resposta. A pergunta pode ser também direcionada a alguém

[1] https://facebook.com
[2] https://plus.google.com/

próximo a um local ou em um prédio específico por meio da indicação do local onde a pessoa se encontra ou por meio de vinculação da postagem com um local do compus. Dessa forma, evidencia-se o foco em redes formadas por colaboração oportunística e não por amizade.

Todavia *groupwares* (software colaborativo) como o Campus Social necessitam de engajamento dos usuários, caso contrário, esse software não alcançará o seu propósito. Por esse motivo entende-se que o software não precisa apenas prover funcionalidades, mas também promover seu uso, pois isso implica no incremento da colaboração. Vale ressaltar Horita et al. (2014) que afirma que a computação colaborativa depende da cooperação entre duas ou mais pessoas para resolver um problema.

Dessa forma, o que se deseja neste trabalho é prover mecanismos e estratégias para atrair e engajar usuários para o Campus Social. Para isso, o presente trabalho propõe o uso de mecanismos de gamificação para aumentar o engajamento dentro do sistema colaborativo Campus Social. Tal proposta observou as características do Campus Social relacionadas a comunicação oportunística e exemplos de outros trabalhos que usaram a gamificação como técnica de engajamento. Para propor uma abordagem adequada de gamificação para o Campus Social, buscou-se analisar as principais técnicas existentes do uso de técnicas de jogos.

O objetivo do presente trabalho é usar adequadamente técnicas de game para aumentar a adesão e o engajamento dos usuários do Campus Social. Adotou-se ranques de usuários, além de incentivos de progressão para aumentar o uso. O ranque também é utilizado para atribuir um grau de confiança às mensagens dos usuários. As mensagens, por sua vez, são marcadas com a contagem de *likes* de *deslikes*. As técnicas utilizadas para aumentar o engajamento devem também promover melhores experiência de uso do software por parte da comunidade como acontece com os games.

O trabalho está organizado da seguinte forma: a seção 2 fundamenta a proposta deste trabalho com a descrição do Campus Social, dos conceitos básicos e dos trabalhos relacionados; a seção 3 descreve a proposta deste trabalho por meio do mapeamento dos mecanismos de gamificação para o Campus Social; e a seção 4 apresenta as considerações finais do trabalho e trabalhos futuros.

2. Fundamentação

Esta seção apresenta os conceitos básicos de gamificação que servirão de base para entendimento da proposta. Também é apresentada uma revisão de literatura com os principais aplicativos para dispositivos móveis e trabalhos acadêmicos que utilizam gamificação dentro do contexto de colaboração e/ou comunicação oportunística.

2.1. O Campus Social

O Campus Social é predecessor do UFRJ Social cuja motivação surgiu a partir da observação da "distância" da informação até os seus interessados. Essa distância representa tanto a separação física dos usuários dentro de um campus universitário como a falta de infraestrutura que proporcione a centralização, organização e manutenção da informação.

O Campus Social se tira proveito da popularização das tecnologias móveis e da comunicação sem fio para proporcionar uma forma de propagação oportunística de informação por meio da colaboração dos frequentadores do campus universitário. A ferramenta oferece mecanismos para os usuários publicarem e comentarem mensagens, as quais são relacionadas a interesses e locais dentro da universidade. Usuários com interesses comuns e com conhecimento sobre o tema de uma mensagem podem colaborar para que outro usuário ou um grupo atinja um interesse. Por exemplo, encontrar onde está sendo realizado um a escola regional de sistema de informação do Rio de Janeiro na Universidade Federal Rural do Rio de Janeiro. Quais prédios e horários de palestras, etc.

Os usuários, ao criarem sua conta, podem informar seus interesses com base em uma lista padrão do sistema. O Campus Social permite que os usuários solicitem a inclusão de novos tópicos de interesse sejam adicionados, contudo novos tópicos precisam ser auditados pelos administradores. Quando uma informação é postada, o criador da mesma pode relacioná-la com os tópicos presentes na lista de interesses. Outra funcionalidade utilizada pelo Campus Social é o uso de localização por prédios no campus o que pode ser feito de duas formas: uso de GPS, ou o usuário que publica a mensagem informa onde está ou se a mensagem está relacionada a um local do Campus. Mensagens são definidas por tópicos e podem ser comendas por qualquer usuário do sistema. A lista de interesses e a localização são utilizadas pelo mecanismo de recomendação de informação utilizado para atualizar a *timeline* dos usuários. A estratégia de recomendação busca possibilitar que usuários com interesses comuns em determinado momento possam colaborar oportunisticamente.

O Campus Social também possui funcionalidades, como mapas, locais pré-cadastrados, entre outros. Descrições mais detalhadas sobre o Campus Social podem ser obtidas em [Oliveira et al. 2012] e [Tabak et al. 2015].

2.2. Gamificação

Deterding et al. (2011) afirmou que gamificação pode ser definida como a utilização de elementos de *gamefulness*, *gameful interaction* e *gameful design* para um propósito específico que se tenha em mente. Em nosso contexto, *gamefulness* refere-se à experiência vivida no uso de um game, *gameful interaction* são todas as ferramentas e objetos e contextos que farão parte da experiência de usuário, e *gameful design* refere-se à prática de elaboração de uma experiência de uso. Dado que o presente trabalho tem o foco as interações (*gameful interaction*) do usuário com o Campus Social, a proposta apresentada foca em características de pontuação para usuários que utilizam esse aplicativo.

São 5 os mecanismos de gamificação *gameful interaction*: (1) os **pontos** os quais são utilizados para mensurar numericamente o avanço dos usuários e são utilizados como meda de troca dentro do sistema por meio de recompensas para usuários; (2) as **insígnias** que servem para representar as recompensas visuais fornecidas aos usuários que progridem (avançam) no sistema; (3) os **níveis** que servem para estimular a progressão e classificar os usuários; (4) os **ranques** para estimular a competição entre os usuários conhecidos (ou com interesse comum no caso do Campus Social); e (5) os **desafios** que buscam aumentar o engajamento por meio da atribuição de tarefas específicas para um usuário ou grupo de usuários.

2.3. Trabalhos Relacionados

Atualmente, existem diversos trabalhos que exploram os benefícios da gamificação como ferramenta de engajamento. Ferramentas envolvendo Sistemas Colaborativos abrangem diversas áreas e junto com esta técnica proporcionam uma melhor experiência de usuário.

No contexto de mobilidade urbana [Waze, 2016] mudou a forma como dirige-se e trafega-se pelos grandes centros urbanos graças as informações em tempo real disponibilizadas durante o trajeto. Esta plataforma fundamente-se na colaboração e recompensa seus usuários com pontuações de acordo com as ações feitas dentro da aplicação.

Na área de ambientes virtuais de aprendizado, [Duolingo, 2016] tornou-se a maior plataforma on-line e gratuita de idiomas graças ao uso de técnicas que permitiram seus usuários aprender múltiplos idiomas paralelamente acompanhando o avanço por meio de ranques, medalhas e personalização. O [Passei Direto, 2016] tornou-se referência ao empregar ranques, níveis e pontos de experiência. Em [Khan Academy, 2016], o uso de pontos, níveis, missões e medalhas permitiu ao usuário realizar diversos cursos e acompanhar seu desenvolvimento de modo interativo.

No contexto humanitário, Horita et al. (2014) usou as técnicas de gamificação junto ao conceito de sistemas colaborativos, com o objetivo de conceber a arquitetura para gerenciar recursos voluntários dentro de um ambiente que sofreu um desastre natural.

No contexto universitário, Walter et al. (2011) apresentou uma aplicação baseada em gamificação cujo objetivo era orientar e guiar o público universitário dentro em questões do seu cotidiano por meio de uma lista de eventos organizada com base no seu nível de graduação do usuário. O CarrerNetwork [Kaplan University, 2016] incorporou o uso da gamificação em sua rede de serviço de carreira (*carrer services network)* visando tornar a busca por um estágio ou emprego mais prazerosa para os estudantes. O presente trabalho diferencia-se dos demais descritos por focar no uso da gamificação como ferramenta de integração e engajamento da comunidade acadêmica como um todo nas questões inerentes a vida universitária dentro do campus.

3. Estratégias de Gamificação para o Campus Social

Foram adotados 4 mecanismos de gamificação para o Campus Social: pontos, insígnias, níveis e ranque. A abordagem do uso desses mecanismos no Campus Social estão descritas nesta seção.

3.1. Pontos

O Campus Social permite que os usuários publiquem diversos tipos de conteúdo (como notificações sobre problemas no campus, eventos dentro da comunidade acadêmica, interesses menos abrangentes como perguntar sobre a localização de uma aula, etc.). Apesar da comunidade colaborar para moderar o conteúdo publicado, percebeu-se que a possibilidade da publicação de conteúdo não verídico (boatos) ou outro uso malicioso poderia causar um impacto negativo sobre a ferramenta diante da percepção da comunidade de um campus universitário. Dessa forma, a abordagem de pontuação busca motivar os usuários a agir da melhor maneira possível. Para tanto, buscou-se modelar o que os autores consideram boas e más práticas de interação diante das funcionalidades para propor uma forma de pontuação.

Dessa maneira, adotou-se os pontos por *expericience* (xp) como ferramenta dentro do sistema. Todas as ações básicas de um usuário irão gerar esse tipo de pontuação. Essas ações representam uso de funcionalidades do sistema para: publicar mensagens; comentar mensagens; receber *likes* nas suas postagens (mensagens ou comentários); publicar eventos; indicar novos usuários em outras mídias; compartilhar informações oriundas do Campus Social em suas mídias sociais; entre outras. As listagem de pontos podem ser visto na Tabela 1.

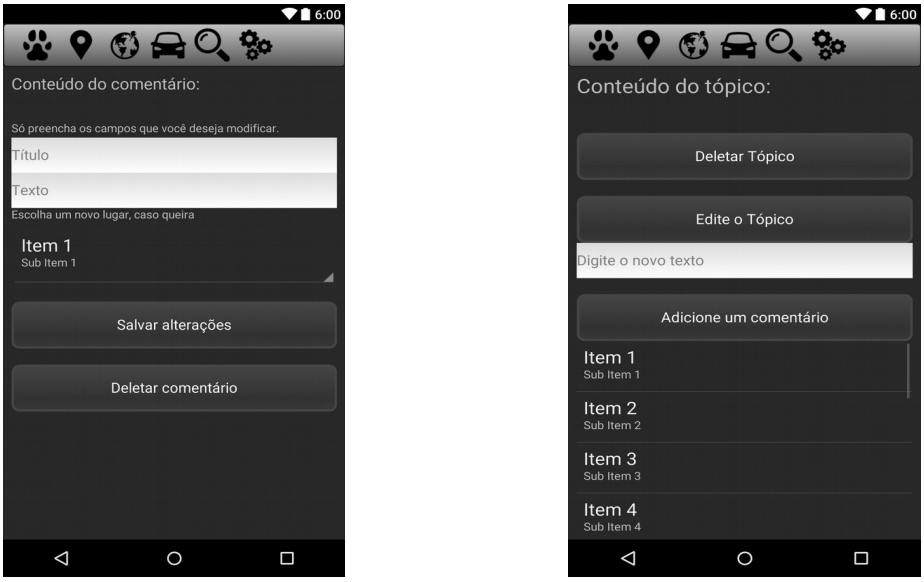
Uma abordagem que permite que os usuários percam pontos também foi adotada. A perda de pontos é a abordagem que desincentiva postagens inadequadas (segundo a definição dos autores) no sistema. Por exemplo, ao postar conteúdo ofensivo (xingando alguém), ou levantando um boato, o sistema saberá que o usuário perdeu 10000000 xp. A despontuação, dependendo da sua motivação, pode até levar ao bloqueio de um perfil até que seja feita a análise da postagem. O Campus Social não possuirá limite máximo de pontuação.

Sobre a despontuação é importante ressaltar que os games no geral também despontuam os jogadores quando esses falham em algum momento ou quando os mesmos realizam ações não permitidas no jogo. Dentro desse cenário, a despontuação nunca foi vista como desmotivante no contexto do *games*, mas foi vista sim como um fator desafiador capaz de agregar emoção aos jogos. Então, se bem utilizada, a possibilidade de perder pontos pode ser vista até como fator motivacional ou emocionador dentro do game. Essa observação faz-se necessária porque no contexto das aplicações com finalidades educacionais a despontuação não é utilizada por ser considerada frustrante nesse contexto. Outra observação é que essas aplicações normalmente fornecem funcionalidades com interações por opções que não possibilitam que os usuários realizem atividades indesejáveis ou mesmo pública. Por exemplo,

responder uma questão objetiva ou completar um sentença cuja resposta não será vista por outros usuários do sistema. A Figura 1 demonstra dois *screenshot* contendo exemplos de ações que podem gerar pontuação.

Tabela 1: Mecanismos de Pontuação

AÇÕES	PONTOS DE XP
Acesso por dia à aplicação	+ 20 xp
Publicar uma mensagem	+ 50 xp
Comentar uma mensagem	+ 50 xp
Avaliar uma mensagem	+ 60 xp
Publicar em outra rede social	+ 100 xp
Enviar e ter um convite aceito	+ 150 xp
30 boas avaliações no post	+ 200 xp
60 boas avaliações no post	+ 500 xp
Mais de 80 boas avaliações	+ 1000 xp
30 más avaliação no post	- 400 xp
60 más avaliações no post	- 600 xp
Mais de 80 más avaliações	- 2000 xp
Publicar conteúdo ofensivo	- 100000000 xp



(a) **(b)**

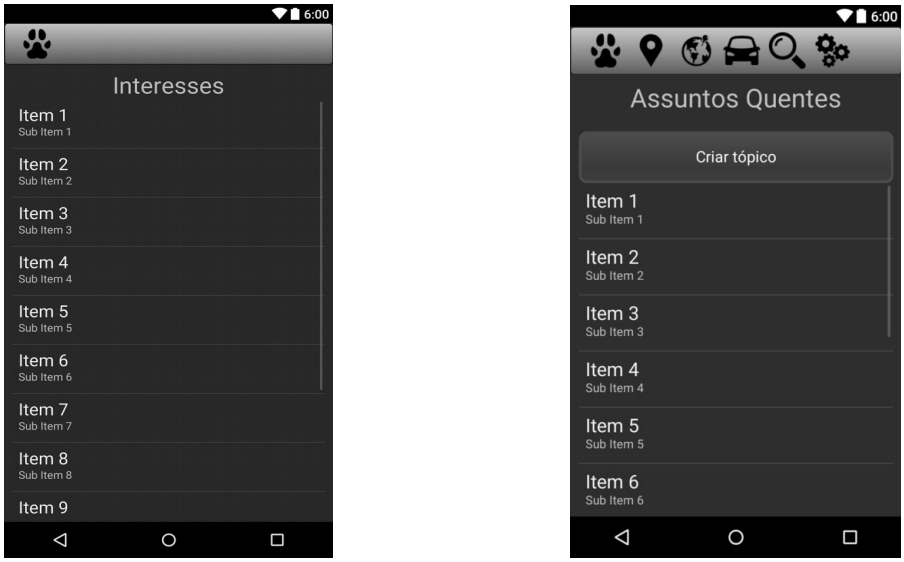
Figura 1:Exemplos de ações padrões que geram pontos de xp. (a) Comentar Post (b) Comentar Tópico.

3.2. Insígnias

O Campus Social agrupa as mensagens por interesse. Tanto os usuários quanto as informações são agrupados por interesse. Os usuários recebem as informações ordenadas por interesse e data. Observando o funcionamento do Campus Social, propõem-se adotar mecanismos de sinalização para diferenciar usuários bem ranqueados no sistema para que essa sinalização possa servir para outros usuários como uma indicação de usuários referências dentro de determinada área de interesse. Dessa maneira, fora feito um modelo de uso de insígnias, todas essas baseadas no sistema de pontuação. A Figura 2 mostra a exibição de tópicos por interesses no aplicativo.

As insígnias, são a representação virtual e visual do avanço conquistado pelo usuário dentro do Campus Social. O objetivo deste item é proporcionar a sensação de premiação por um

serviço prestado à comunidade acadêmica e por reconhecimento por outros usuários da experiência de um usuário. Do ponto de vista dos usuários que buscam informação, estes podem entender as insígnias como indicativo da autoridade que um certo usuário possui sobre um tema.



(a) **(b)**

Figura 2: Painel de tópicos por interesses do Campus Social. (a) Lista de Interesses e (b) assuntos por Interesses.

Foram definidas 4 insígnias para os usuários do Campus Social: bronze, prata, ouro e platina (com diferenciações da insígnia por áreas de interesse). A Tabela 2 mostra como como estão sendo atribuídas as insígnias dos usuários. Pode-se perceber que as mesmas são atribuídas pelo engajamento dos usuários do sistema.

Tabela 2: Exemplo de insígnia de engajamento

INSÍGNIAS	Insígnias de Bronze	Insígnias de Prata	Insígnias de Ouro	Insígnias de Platina	Insígnias de Diamante
REQUISITOS	20 comentários bem avaliados em uma semana	40 comentários bem avaliados em uma semana	60 comentários bem avaliados em uma semana	80 comentários bem avaliados em uma semana	100 comentários bem avaliados em uma semana

3.3. Níveis

O uso de níveis dentro do Campus Social têm por objetivo refletir a evolução geral do usuário dentro do uso da aplicação. Vale destacar que diferente das insígnias, os níveis não estão ligados a ações específicas dentro sistema, mas sim com a quantidade de pontos de experiência que são acumulados no dia a dia dos usuários. Foram definidos 7 níveis, todos com nomes pertencentes ao universo universitário. A Tabela 3 mostra o exemplo de cada nível e suas faixas de pontuação. No exemplo utilizou-se termos conhecidos no contexto da UFRRJ e o termo genérico no contexto das universidades brasileiras.

3.4 Ranque

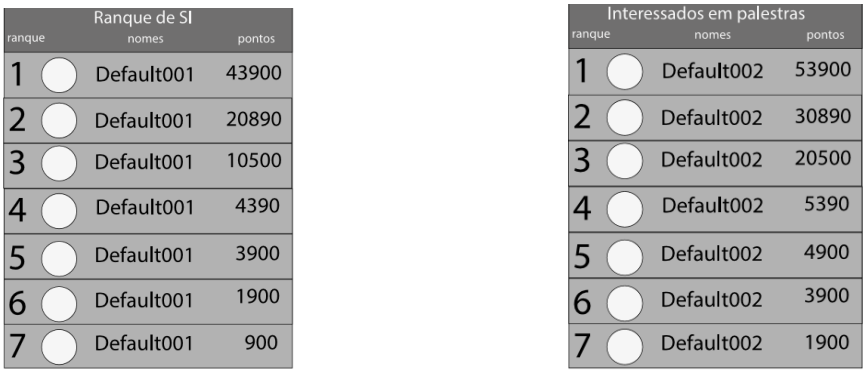
A aplicação do mecanismo de ranqueamento utiliza a distribuição de tópicos em interesses presente no Campus Social como forma de organização dos ranques. Buscou-se modelar esse mecanismo de forma a proporcionar que o usuário se sinta desafiado a concorrer com pessoas próximas ou com usuário com os mesmos interesses. Segundo Klock et al. (2014) um ranque

tem por objetivo comparar jogadores/usuários do sistema e gerar um senso de competição entre os mesmos. Por esse motivo e observando a proximidade dos usuários de um campus universitário, esse mecanismo foi projeto com abordagem descrita.

A Figura 3 apresenta um exemplo de ranques no Campus Social. A Figura 3(a) apresenta um ranque dos usuários por curso. A Figura 3(b) apresenta o ranque por um interesse dos que são pré-definidos no sistema.

Tabela 3: Níveis e suas respectivas faixas de pontuação

Bixo (Calouro) 0 ~ 100 xp	Iniciante dentro do sistema, assim como um calouro na Faculdade.
Bixo' (Calouro') 101 ~ 400 xp	Usuário com experiência básica , mas ainda há um longo caminho à percorrer
Veterano 401 ~ 1000 xp	Assim como em uma faculdade, veterano significa um usuário experiente no uso da plataforma
Vôterano 1001 ~ 2500 xp	Termo usado na UFRRJ para alunos com 8 períodos cursados, nesse caso um usuário muito experiente
Lenda 2501 ~ 5000 xp	Assim como o nome sugere, nesse nível o usuário possui uma considerável quantidade de pontos
Mestre 5001 ~ 10000 xp	Fazendo referência à alunos de mestrado, nesse nível a pessoa está bem acima de típicos usuários
Mito 10001 xp ~∞	Nível máximo a ser alcançado. Representa o auge de um usuário dentro da plataforma



(a) **(b)**

Figura 3: Painel de ranques do Campus Social. (a) Ranque por Curso e (b) ranque por Interesses.

4. Considerações Finais

Este trabalho apresentou a descrição da abordagem de gamificação para adesão e engajamento de usuários para o Campus Social [Tabak et al. 2015], bem como conceituação desses usuários dentro do sistema. Trata-se de um trabalho em desenvolvimento que acrescenta técnicas de game para um aplicativo de comunicação oportunística para campus universitário. Foi realizada uma descrição da proposta e uma análise dos elementos que juntos compõem o conceito de gamificar um sistema. Concomitantemente, foram apresentadas propostas de utilização da gamificação dentro do sistema em desenvolvimento Campus Social, sendo o uso cada uma delas justificado. A proposta foi desenvolvida com base em outras aplicações disponíveis para dispositivos móveis e no levantamento de trabalhos relacionados.

Apesar de se tratar de um trabalho em desenvolvimento, percebeu-se, com base na literatura relacionada, o potencial do uso de gamificação para os objetivos desejados (adesão, engajamento e diferenciação dos usuários para atribuí-lhes uma indicação de expertise para os

interesses da comunidade de um campus). A abordagem proposta do uso de gamificação está em fase de implantação no Campus Social.

Como trabalho futuro pretende-se: concluir a implementação da proposta apresentada; desenvolver novas soluções que contemplem melhoras no que diz respeito à *gamefulness* (ou seja tornar mais agradável a experiência do usuário por meio da gamificação); elaborar e implementar uma abordagem de *gameful design*. Após a nova abordagem entrar em funcionamento, espera-se verificar se o uso da ferramenta aumentou tanto do ponto de vista da adesão de novos usuários quanto do ponto de vista da colaboração dos usuários do sistema. Por fim, pretende-se verificar a taxa de permanência, o perfil dos usuários (faixa etária, o que faz no campus, etc.) e as funcionalidades mais populares.

Referencias

Deterding et al. (2011) “Situated motivational affordances of game elements: a conceptual model.”, In: Gamification: Using Game Design Elements in Non-Gaming Contexts, a Workshop at CHI. Presented at CHI 2011. ACM, Vancouver, BC, pp. 1–4.

Duolingo. (2016) “Duolingo”, Disponível em: < https://pt.duolingo.com/>. Acesso em 16 de Outubro de 2016.

Facebook. (2016) Site Oficial, Disponível em: <https://pt-br.facebook.com/>. Acesso em 16 de Outubro de 2016.

Google. (2016) “Google Plus”, Disponível em: <https://plus.google.com/>. Acesso em 16 de Outubro de 2016.

Horita et al. (2014) “A Gamification-based Social Collaborative Architecture to increase resilience against natural disasters”, In: Simpósio Brasileiro de Sistemas de Informação (SBSI), Londrina, Brazil.

Kaplan University (2016) “CarrerNetwork”, Disponível em: <https://www.universitybusiness.com/news/kaplan-incorporates-gamification-its-career-services-network>. Acesso em 16 de Outubro de 2016.

Khan Academy (2016) “Khan Academy”, Disponível em: <https://pt.khanacademy.org/>. Acesso em 16 de Outubro de 2016.

Klock et al. (2014) “Análise das técnicas de Gamificação em Ambientes Virtuais de Aprendizagem”, In: CINTED- Novas Tecnologias na Educação.

Oliveira et al. (2012) “Propagação Colaborativa e Recomendação De Informações Utilizando Computação Móvel E Dados Georreferenciados.” In: SimSocial - Simpósio em Tecnologias Digitais e Sociabilidade, 2012, Salvador. Anais do Simpósio em Tecnologias Digitais e Sociabilidade, 2012.

Passei Direto. (2016) Site Oficial, Disponível em: <https://www.passeidireto.com/>. Acesso em 16 de Outubro de 2016

Tabak et al. (2015) “Campus Social: uma ferramenta para trocas oportunísticas de informações em campi universitários." In: 42º Seminário Integrado de Software e Hardware (SEMISH), Recife”

Walter et al. (2011) “Orientation passport : using gamification to engage university students”. In Proceedings of the 23rd Australian Computer-Human Interaction Conference, ACM, Australian National University, Canberra, ACT.

Waze. (2016) “Waze Mobile”, Disponível em: <https://www.waze.com/pt-BR/>. Acessado em 15 de Setembro de 2016.

Sistomate: Sistema Inteligente de Suporte à Decisão no Auxílio ao Combate da Requeima em Culturas de Tomate

Gustavo S. Oliveira[1], Gizelle K. Vianna[1]

[1]Departamento de Matemática, Universidade Federal Rural do Rio de Janeiro (UFRRJ), Seropédica, Rio de Janeiro, Brasil.

{gustavo1071af,gkupac}@ufrrj.br

Abstract. *Nowadays farmers have realized that the use of intelligent and decision support systems in agriculture is more than just a trend, it is an act of survival. Their adoption is justified by the growing needs of consumer market and due to food safety issues, respect for environment, and health of labours. However, few systems fulfill the needs of small Brazilian tomatoes producers. This paper presents an approach focused on improving the quality of the tomato crops. Thus, to achieve such goal, we designed and implemented a computational environment to support decision-making, called Sistomate, related to the tomato production.*

Resumo. *Cada vez mais, os agricultores têm percebido que a tomada de decisões e uso de sistemas inteligentes é mais que uma simples tendência, mas uma ação de sobrevivência e obrigação, justificada pela exigência cada vez maior do mercado consumidor quanto às questões de segurança alimentar, respeito ao meio ambiente e saúde do trabalhador rural. No entanto, existem poucos sistemas que atendem aos pequenos produtores de tomate. Este trabalho apresenta uma abordagem focada na melhoria da qualidade das culturas de tomate e, para atender a esse objetivo, desenvolvemos um ambiente computacional, chamado Sistomate, de apoio à tomada de decisões relacionadas à produção de tomates.*

1. INTRODUÇÃO

O tomate é um produto importante na economia agrícola do Brasil, um dos países mais bem colocados na produção do fruto. A safra brasileira de tomate foi estimada, em 2014, em mais de 1,9 milhão de toneladas, colocando o país como o 7º produtor mundial de tomates para processamento [RABELO 2015]. Entretanto, devido ao grande número de pragas e doenças que afetam os tomateiros, o tomate, juntamente com a batata, se destaca pela grande quantidade de agrotóxicos utilizada nas plantações [NORTERS 2015].

No meio das doenças que afetam o tomateiro no Brasil e no mundo, a requeima, a causada pelo oomiceto *Phytophthora infestans,* vem a ser uma das mais destrutivas, podendo comprometer toda a área de produção em poucos dias [LOPES 1994; STEVENSON 1983]. Algumas condições e fatores como o clima, localização da área plantada, modo de implantação e de condução da lavoura são determinantes para a frequência e intensidade das doenças que afetam o tomate [FILGUEIRA 2008]. O uso de

agroquímicos nos tomateiros de forma abusiva pode causar problemas severos ao meio ambiente e à saúde dos produtores e consumidores [CORREA 2009].

Este trabalho descreve a fase final de desenvolvimento de um sistema de apoio à decisão intitulado Sistomate, que possui recursos para prever a evolução da requeima em cultivos de tomate. O objetivo do projeto é gerar alertas de ocorrência da requeima e simulações de cenários de propagação da contaminação e ainda propor alternativas para o combate à doença, com o auxílio de dados meteorológicos e modelos de previsão da requeima. Todo o sistema de decisão está baseado em ferramentas de classificação automáticas das alterações foliares nos tomateiros de uma determinada propriedade rural, desenvolvidos em trabalhos anteriores [VIANNA 2013a; 2013b; 2014].

2. MATERIAL E MÉTODOS

2.1. Método de classificação das amostras das folhas de tomate

Para a tarefa de identificação das áreas foliares atingidas pela doença foram utilizadas imagens reais das plantas submetidas a um filtro de cor vermelho/verde criado por [VIANNA 2013a; 2013b; 2014]. O procedimento da filtragem ocorre de forma automatizada, considerando o princípio de que a cor da planta saudável é o verde intenso, tons amarelados ou marrons indicam a ocorrência de alguma lesão e tons diferentes fazem parte do fundo da imagem e devem ser ignorados. O procedimento analisa cada imagem em três etapas: Redução Proporcional, Filtragem e Processamento e Armazenamento (Figura 1).

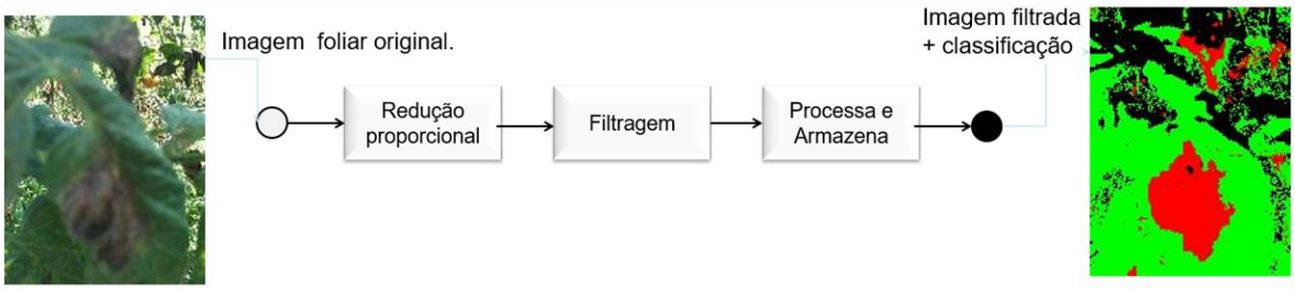


Figura 1. Processo de classificação das amostras de tomates.

Tabela 1 – Tons de cor atribuído a cada exemplar de tomateiro, após o processo de filtragem das imagens coletadas no campo.

Estado	0	1	2	3	4	5	6
% de infecção	≈ 0	≈ 3	≈ 12	≈ 22	≈ 40	≈ 60	≈ 77
Tom de cor do tomateiro no mapa	Verde Escuro	Verde	Verde Claro	Amarelo	Laranja	Laranja Escuro	Laranja Vermelho

Para cada imagem já filtrada, o processamento se inicia com a contagem dos *pixels* que correspondem às cores Vermelha, Verde e Preta e termina com o cálculo da razão entre os pixels vermelhos e verdes, correspondendo ao percentual de infecção pela requeima. Esse percentual de infecção classifica cada amostra dentro de uma das faixas descrita na tabela 1, conforme as chaves de classificação propostas por [CORREA 2009], em uma escala logarítmica de valores. Por fim, essa classificação individual é armazenada

no banco de dados do sistema, juntamente com um conjunto de dados auxiliares calculados durante o processamento de filtragem.

2.2. Modelo de Previsão Empregado no Sistema

De acordo com Integrated Pest Management Program of California University – UC IPM (2016) existem vários modelos de previsão bem avaliados da requeima em cultivos de tomates e batatas. A maioria dos métodos descritos exigia a inserção de dados meteorológicos de que não dispúnhamos, como por exemplo, o Wallin (1962) que necessita da informação sobre o número de horas do dia com umidade maior ou igual a 90%. Portanto, o método escolhido foi o modelo de previsão Hyre (1954). De acordo com o autor, um surto inicial de requeima é previsto de 7-14 dias após 10 dias favoráveis consecutivos para a requeima. Um dia favorável ocorre quando a média das temperaturas dos 5 dias anteriores estiverem abaixo de 25,5 ºC (dias com temperatura mínima menor ou igual a 7,2 ºC não são contados), e 10 dias anteriores com um total de precipitação igual, ou maior, que a média do total de precipitação durante a mesma quantidade de dias na temporada, 1.21 polegadas (30.734 milímetros).

2.3. Obtenção dos Dados Meteorológicos

O clima é fundamental para o desenvolvimento, ou não, da requeima nos campos de tomate. Todavia, descobrir a previsão de dados meteorológicos não é uma tarefa trivial se considerarmos que o sistema proposto fará simulações de vários dias, mais do que a janela de previsões que pode se obter diretamente de portais meteorológicos online. Além disso, o modelo de previsão criado utiliza dados históricos de uma região e esses foram obtidos por meio do portal do Instituto Nacional de Meteorologia – INMET (2016).

No estudo de caso apresentado neste artigo, foram utilizados os seguintes dados da cidade de Paty de Alferes, RJ: temperatura média, umidade relativa do ar média, temperatura mínima, temperatura máxima e precipitação do período. Essa cidade foi escolhida para a simulação por ser o município maior produtor de tomates no Estado do Rio de Janeiro e foram coletados os dados desde 01/01/1999 até 01/01/2015.

O sistema terá a opção de escolher qual o tamanho da janela de dados que será usada no cálculo da média histórica. Essa média histórica é usada para criar uma estimativa dos dados meteorológicos para os períodos simulados, conforme o exemplo da Tabela 2, necessários para o modelo de previsão de Hyre (1954).

Tabela 2 – Exemplo de construção da média histórica da variável temperatura, utilizando uma janela de 5 anos. Esse valor será usado como temperatura esperada para o próximo dia 05 de junho de 2016.

Uma Iteração da Simulação	Temp. Média do mesmo dia e mês da Iteração em 2010	Temp. Média do mesmo dia e mês da Iteração em 2011	Temp. Média do mesmo dia e mês da Iteração em 2012	Temp. Média do mesmo dia e mês da Iteração em 2013	Temp. Média do mesmo dia e mês da Iteração em 2014	Resultado da Temperatura média da Interação
05/06/16	17,42	11,24	21,64	18,48	16,88	**17,13**

2.4. Descrição do Autômato Celular

De acordo com os estudos feitos e a visando simular a realidade, utilizamos um Autômato Celular para modelar a dinâmica da requeima. O autômato celular foi definido no domínio

bidimensional, com vizinhança de Moore e dinâmica definida por uma função de transição probabilística. A construção da função de transição foi embasada pelo modelo de previsão Hyre (1954) de forma a definir os dias favoráveis para requeima e quando se inicia e o surto da mesma. A classe do autômato utilizado neste trabalho baseou-se em uma reengenharia do Jogo da Vida (Game Of Life) [CONWAY 1970].

A matriz do autômato representa uma área cultivada, ou um talhão, de uma produção de tomate e, a fim de adequá-la ao nosso domínio, as colunas correspondem às linhas da área cultivada. Dentro de cada coluna da matriz, encontram-se as plantas de tomate, dispostas em linhas. Coerentemente, cada célula da matriz representará uma planta de tomate que possuirá um valor de estado associado a ela (Figura 2).

O autômato será inicializado com dados preenchidos pelo usuário do software, como o *tamanho da janela* de dados históricos e a *direção do vento*. A variável *direção do vento* controla o sentido das alterações de estado das células do autômato. As mudanças de estado dos vizinhos de uma célula qualquer, que representa as infecções da vizinhança causada por uma célula infectada, acontecerão apenas na célula vizinha referente a direção do vento e as células vizinhas ao lado da célula referente a direção do vento, como mostra a Figura 3.

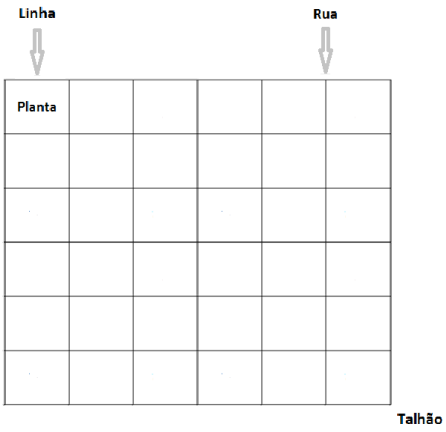


***Figura 2*. Matriz do autômato baseada nos conceitos de manejo de tomates.**

Semelhante ao autômato de Conway, em que a célula x tem o seu estado alterado na próxima iteração dependendo do estado das células vizinhas, o próximo estado da célula c(i,j), onde i é a linha e j é a coluna, que chamaremos de $E'(c(i,j))$, depende do estado atual de c(i,j), que chamaremos de $E\big(c(i,j)\big)$ e dos estados de seus vizinhos, em uma vizinhança de tamanho 8. Além da possibilidade de uma célula infectada piorar o estado da célula atual, ele também pode melhorar, ou pelo menos o dano ser amenizado, caso uma forma de combate à doença, que chamaremos de C, esteja sendo utilizada. Cada vizinho afeta a célula c(i,j) de forma ponderada, segundo a Tabela 3, definida de acordo com os fatores indicados por Hyre (1954). A influência ponderada de cada vizinho se baseia na quantidade de surtos Qs, quantidade de dias favoráveis à requeima Qd e o estado atual E da célula c(i,j), sendo calculado seguindo as regras da Tabela 3.

Ao mesmo tempo, os valores das células que compõem essa vizinhança irão alterar seu valor no próximo passo, formando a nova matriz de estados. Existem duas formas de combate sendo testadas aqui e, de acordo com a literatura, sua eficiência é tal que, no combate tipo 1, o estado será reduzido em 30% do estado atual e, no combate tipo

2, em 20%. A dinâmica do autômato pode ser resumida pela fórmula 1 que define a função de transição do autômato e pela Tabela 3, abaixo:

$$E'\big(c(i,j)\big) = E(c(i,j) + \sum_{n=1}^{8} P\left(v_n\big(c(i,j)\big)\right) - C * E(c(i,j)) \quad (1)$$

Tabela 3 – Regras para o cálculo do Peso P.

		$E(c_{(i)})$					
		1	**2**	**3**	**4**	**5**	**6**
Qs > 1	**Qd >= 10**	0.1	0.8	1.4	1.6	1.8	2
	10 > Qd >= 7	0.1	0.5	1	1.1	1.2	1.4
Qs > 3	**7 > Qd >= 5**	0.2	0.4	0.6	0.7	0.8	0.9

As regras que controlam a dinâmica da simulação, alterando os estados das células de modo a representar o espalhamento da requeima no mapa, ou matriz. São definidas por um conjunto de parâmetros, cujos valores foram baseados e ajustados de acordo com o modelo de previsão de Hyre (1954). Os ajustes foram importantes para aprimorar e aproximar o resultado da simulação com a realidade dos campos de tomates afetados pela doença.

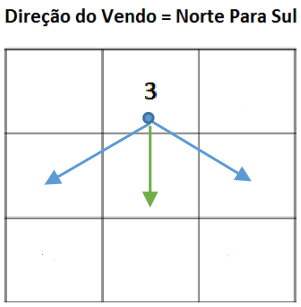


***Figura 3*. Exemplo de uma célula de estado 3 infectando seus vizinhos de acordo com a direção do vento. Note que a seta em verde representa a célula vizinha referente a direção do vento.**

3. RESULTADOS E DISCUSSÃO

O protótipo do sistema móvel foi desenvolvido em linguagem Java e já é capaz de mapear as ruas e linhas de um talhão, registrar safras e detectar tomateiros infectados por requeima de uma propriedade agrícola e simular cenários de espalhamento da requeima em um período determinado com a possibilidade de parar a simulação e escolher um método de combate da doença e prosseguir a simulação. As principais funções do sistema proposto são o módulo de processamento e classificação das imagens digitais e o simulador de cenários de propagação da contaminação e de alternativas para combate à doença.

No módulo processamento e classificação das imagens digitais, as imagens são classificadas dentro da escala de cores já descrita e ilustradas sob a forma de uma meta-

representação do tipo grid que considera as informações de georeferenciamento das imagens juntamente com os dados de cada tomateiro mapeado pelo usuário do sistema. O mapa assim construído representa conceitualmente a plantação que está sendo monitorada pelo conjunto de softwares (figura 4).

No mapa conceitual (Figura 3a), é possível selecionar qualquer uma célula do grid e recuperar as informações da amostra correspondente, inclusive a imagem original da folha, o estado atual da planta e o endereçamento da planta no talhão (Figura 3b).

Por outro lado, o módulo da simulação de cenários de propagação da contaminação e de alternativas para combate da doença foi desenvolvido usando autômato celular como detalhado anteriormente. Neste módulo é possível fazer simulações do espalhamento da requeima no mapa representativo da área cultivada e criar estratégias de combate da doença (Figura 5). A simulação é interativa e simples, onde o usuário pode pausar, continuar e reiniciar a simulação a qualquer fase da simulação corrente.

a) **b)**

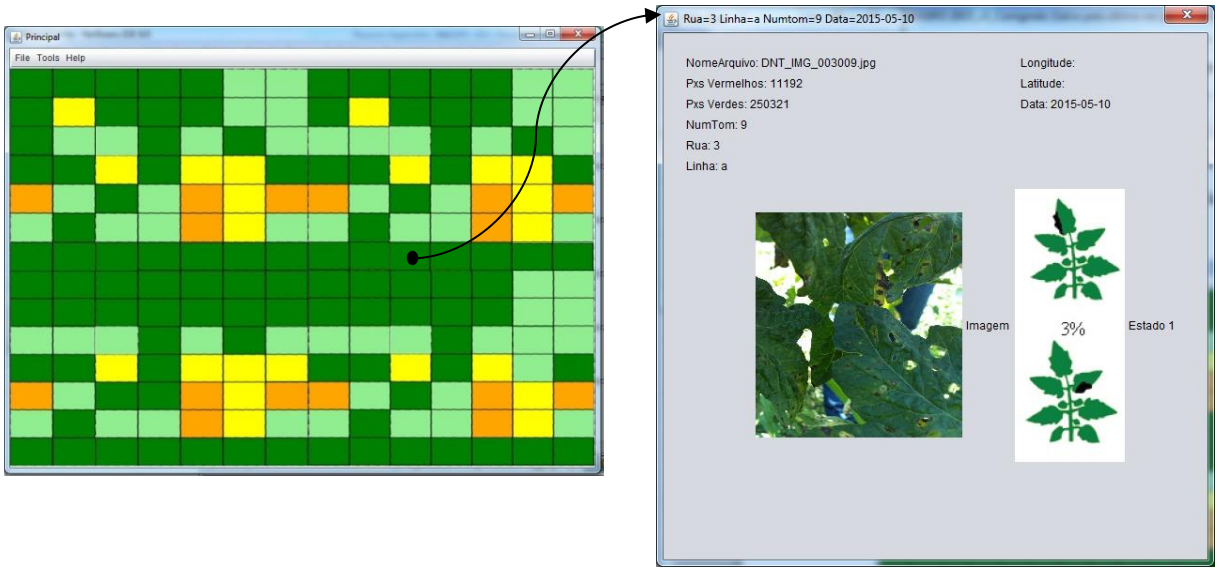


Figura 4. (a) Mapa conceitual de tomateiros de um talhão monitorado pela ferramenta contendo amostras de teste. (b) Exemplo de um tomateiro selecionado no mapa.

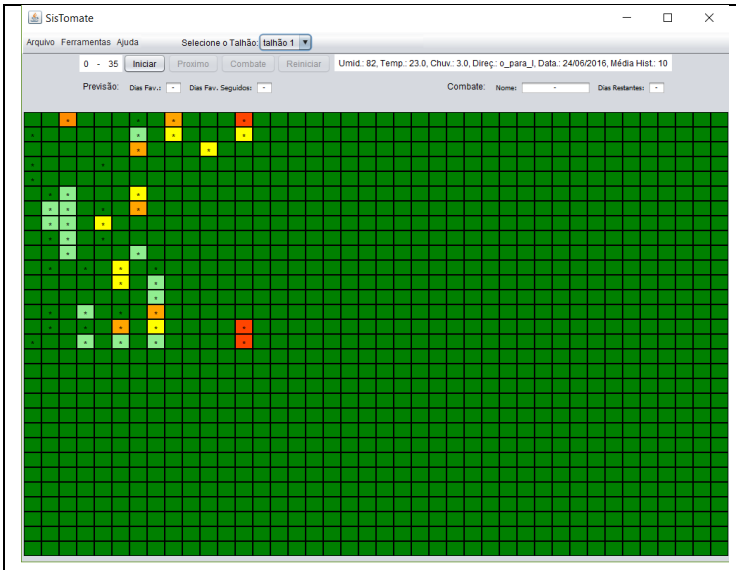
O Sistomate também permite o uso de combates durante a simulação, que pode resulta em um novo rumo para a simulação, diminuindo os estados dos tomateiros do talhão dependendo do nível de contaminação do talhão, dos fatores climáticos e da forma de atuação do combate escolhido.

2. CONCLUSÕES

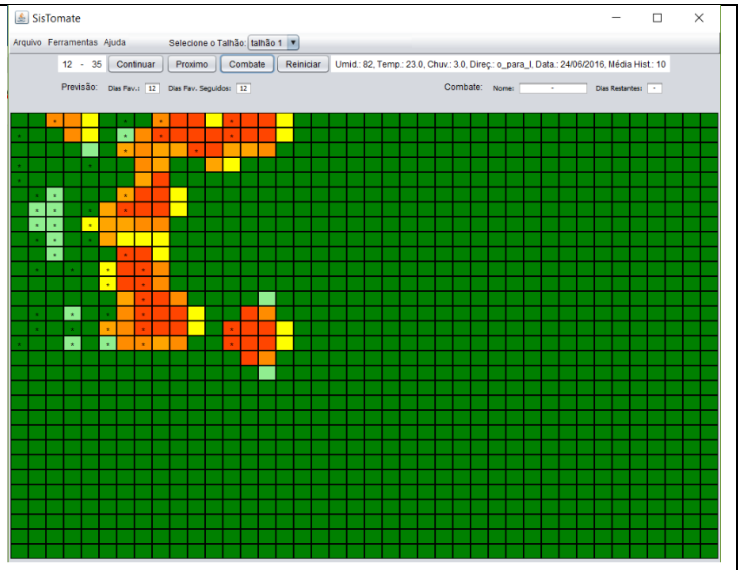
Este trabalho apresentou uma abordagem computacional, com um protótipo em fase já em fase de testes, que já produziu resultados que podem ser empregados em pequenas propriedades onde se pratica a agricultura familiar para o auxílio em tomadas de decisão. A meta é desenvolver ferramentas baseadas em software livre e distribuir as mesmas pelos

produtores do Estado do Rio de Janeiro, contribuindo para o monitoramento da requeima, hoje realizado de forma manual pelo produtor, aumentando a produtividade dessas lavouras.

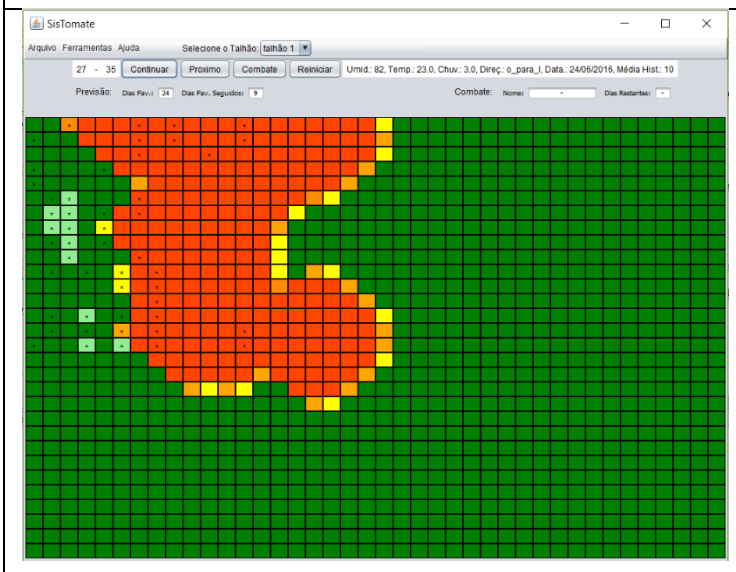
Para trabalhos futuros, será criado um painel de estatísticas após uma simulação, mostrando o desempenho da estratégia feita na simulação com números e gráficos. Também será incluído no projeto um sistema de alertas de possíveis surtos da requeima para produtores vizinhos.



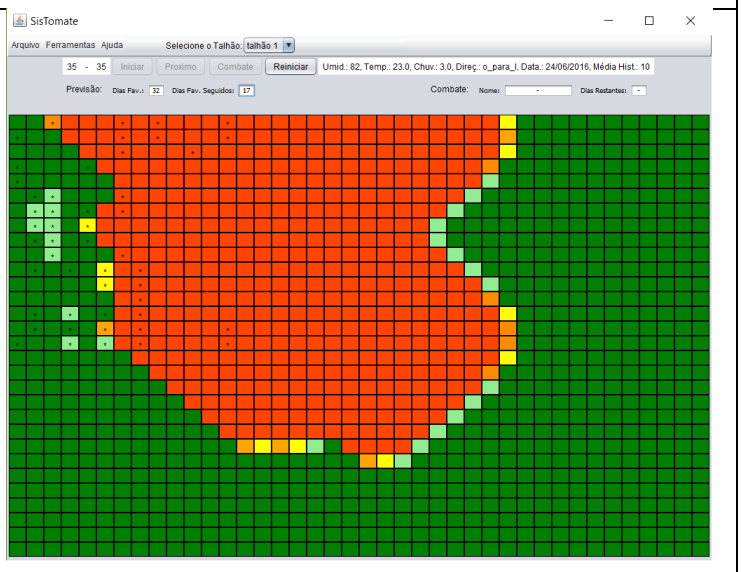
a) **Simulação não iniciada com os estados de acordo com a coleta *in loco (células com '*')***



b) **Representação do estado de cada tomateiro na iteração 12 da simulação.**



c) **Representação do estado de cada tomateiro na iteração 27 da simulação.**



d) **Representação do estado de cada tomateiro na iteração 35 (final) da simulação.**

Figura 5. Simulação sem uso de combate feita para a data 24/06/2016, com a direção do vento Oeste para Leste e 35 iterações sobre um talhão de testes com 1200 elementos.

Referências

CONWAY, J. (1970) “The game of life. Scientific American”, 223(4), p.4.

CORREA, F.M., BUENO FILHO, J.S.S., and CARMO, M.G.F. (2009) “Comparison of Three Diagrammatic Keys for the Quantification of Late Blight in Tomato Leaves”, Plant Pathology 58, p.1128-1133.

FILGUEIRA, F. A. R. (2008) “O novo manual de olericultura”. 3. ed. Viçosa: Editora da UFV.

HYRE, R.A. (1954) “Progress in forecasting late blight of potato and tomato”. Plant Disease Reporter, Illinois, v. 38, n. 4, p. 245 - 253.

INTEGRATED PEST MANAGEMENT PROGRAM OF CALIFORNIA UNIVERSITY– UC IPM. (2016) http://www.ipm.ucdavis.edu/DISEASE/DATABASE/potatolateblight.html. Junho.

LOPES, C.A.; Santos, J.R.M. (1994) “Doenças do tomateiro”. Brasília: EMPRAPA/CNPH. 67 p.

NORTERS. (2015) “Uso de agrotóxicos preocupa especialistas em Meio Ambiente”. http://www.norters.com.br/site/?page=post&id=25409-Uso-de-agrot%C3%B3xicos-preocupa-especialistas-em-Meio-Ambiente, Junho.

VIANNA, G. K., & CRUZ, S. M. S. (2013a) “Análise Inteligente de Imagens Digitais no Monitoramento da Requeima em Tomateiros”. Anais do IX Congresso Brasileiro de Agroinformática. Cuiabá, MT, Brazi.

VIANNA, G. K., & CRUZ, S.M.S. (2013b) “Redes Neurais Artificiais aplicadas ao Monitoramento da Requeima em Tomateiros”. Anais do X Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), Fortaleza, CE, Brazil.

VIANNA, G. K. ; CRUZ, S.M.S. (2014) Using Multilayer Perceptron Networks in Early Detection of Late Blight Disease in Tomato Leaves. In: ICAI´14, 2014, Las Vegas. Proceedings of the 16th International Conference on Artificial Intelligence.

RABELO, M. (2015) “Faeg participa do Congresso Brasileiro de Tomate Industrial”. http://sistemafaeg.com.br/noticias/10796-faeg-participa-do-congresso-brasileiro-de-tomate-industrial. Maio.

STEVENSON, W.R. (1983) “An integrated program for managing potato late blight”. Plant Disease, St. Paul, v.67, n.9, pages 1047-1048.

WALLIN, J.R. (1962) “Summary of recent progress in predicting late blight epidemics” in United States and Canada. American Potato Journal, Orono, v. 39, n.3, pages 306 - 312.

SigaCiente: Uma ferramenta para inferência do trânsito e de rotas seguras baseada em dados sociais

Thamiris Martins Secron[1], Eliel Roger da Silva[1], Claudio Miceli de Farias[2], Tiago Cruz de França[1]

[1]DEMAT – Universidade Federal Rural do Rio de Janeiro (UFRRJ)

[2]INCE – Instituto Tércio Pacitti (UFRJ)

thami_secron@yahoo.com.br, {elielrogernic,cmicelifarias}@gmail.com, tcruzfranca@ufrrj.br

Abstract. *Mobility Crisis and lack of security are known problems in big cities as the largest Brazilian cities. This paper presents a Geographic Information System which uses social data for identifying traffic occurrences and collecting crime data to help the population in planning of their daily commute. In this way, we attempted to use Crowdsourcing data to alert recurring incidents on the roads through monitoring, analysis and visualization of traffic conditions. In addition, the tool allows the recommendation of safe routes based on the crime rate in a given region.*

Resumo. *A crise na mobilidade urbana e a falta de segurança são problemas conhecidos pelas grandes metrópoles brasileiras. O presente trabalho apresenta um Sistema de Informação Geográfica que faz uso de dados sociais para identificar ocorrências do trânsito e coletar dados criminais a fim de ajudar a população no planejamento do seu deslocamento diário. Com dados de crowdsourcing foi possível observar o potencial de melhoria da experiência dos usuários ao utilizar sistemas de recomendação de rota sob o ponto de vista do aumento de informações como também da indicação do grau de risco da rota baseado na criminalidade de pontos no trajeto.*

1. Introdução

A Região Metropolitana do Rio de Janeiro, assim como a maioria das grandes cidades do país, sofre diariamente com o trânsito caótico. A falta de investimentos em transporte coletivos e o fácil acesso ao crédito nos dias atuais incentivam o uso do transporte individual. Concomitante a esses eventos, tem-se o crescente uso de mídias sociais e sistemas de localização como *Google maps*[1] e *Waze*[2] como tentativa para inferir qual a rota mais rápida, obter informações sobre o trajeto e o tempo estimado no caminho a percorrer [Silva *et al* 2015].

Todavia relatos de situações de perigo passados por usuários de GPS têm acontecido. As pessoas, ao buscarem rotas alternativas, são levados a "áreas de risco" – por exemplo, comunidades carentes do Rio de Janeiro – e se deparam com cenas de violência como roubos de veículos, sequestros e assassinatos [Darlington 2015]. Isso tem repercutido negativamente na imprensa internacional que faz um alerta sobre questões de segurança e o uso de aplicativos GPS em estados que tem bairros controlados pelo crime [Darlington 2015].

Além disso, os sistemas de recomendação de rotas geralmente não fornecem informações sobre o trajeto recomendado que aproximem a rota do mapa cognitivo dos

1 https://maps.google.com
2 https://www.waze.com

usuários. Mapas cognitivos representam o processo pelo qual um organismo representa o ambiente em seu próprio cérebro [Laszlo *et al.* 1995 *apud* Bastos 2002]. Esses sistemas levam em consideração apenas o tempo e a distância do trajeto. Especialmente trafegando por rotas desconhecidas, os usuários podem sentir-se frustrados ao não suprir as expectativas sobre os trajetos percorridos (reconhecimento de pontos, ruas, qualidade das vias etc.) [Silva *et al.* 2015].

O presente trabalho apresenta o SigaCiente, um aplicativo que recomenda rota utilizando dados sociais (extraídos de mídias sociais *online* e obtidos com a colaboração dos usuários) para recomendar rotas com base em dados de segurança (por exemplo alto índice de acidentes, frequência de assaltos, entre outros) e agregar informação às mesmas. O objetivo é considerar os riscos que os mesmos são submetidos ao seguir um caminho sem prévio conhecimento do trajeto bem como melhorar a experiência dos usuários de sistemas de recomendação de rotas.

O trabalho está organizado da seguinte forma: a seção 2 apresenta os trabalhos relacionados ao tema abordado e contextualiza o leitor no estado da arte; a seção 3 apresenta a descrição do SigaCiente e provê uma visualização conceitual do sistema proposto; a seção 4 descreve o protótipo desenvolvido e cita as abordagens e tecnologias utilizadas; e por fim a seção 5 apresenta as considerações finais e levanta os trabalhos futuros.

2. Trabalhos Relacionados

Atualmente, existem trabalhos que exploram a rápida propagação da informação nas mídias sociais. Ferramentas envolvendo Sistemas de informação Geográfica e *crowdsourcing* abrangem diversas áreas e disseminam dados de forma mais interativa e democrática. Na área da saúde, o *HealthMaps* tornou-se líder global para acompanhamento de surtos de doenças e epidemias fazendo uso de fontes de notícias, discussões de especialistas e relatos de testemunhas. O *Onde Fui Roubado* é uma ferramenta que através de informações da população faz um mapa criminal de diversas cidades brasileiras.

No contexto do trânsito, Lima *et al.* (2012) analisou dados contextuais dinâmicos do trânsito oriundos de mídias sociais e informações contextuais dinâmicas do trânsito e apresenta rotas recomendadas para o usuário. Silva *et al.* (2015) avaliaram os aplicativos *Google Maps* e *Waze* (segundo os autores, os mais populares) quanto a recomendação de rota fornecida por estes e verificaram a coerência dos aplicativos com relação aos mapas cognitivos dos usuários. Como resultado, os autores concluíram que esses sistemas de recomendação de rotas utilizam pouco os elementos comuns à construção de caminhos feitos por humanos, e dessa maneira, as rotas recomendadas diferem bastante daquelas seguidas pelos humanos que conhecem o trajeto principalmente por causa da forma como funcionam os sistemas de recomendação desses aplicativos. O presente trabalho diferencia-se dos demais descrito por focar na recomendação de rota considerando questões de segurança e na inclusão de informações sobre o que está ocorrendo no trajeto recomendado.

Ainda no contexto de mídias sociais, Franco (2013) e Endarnoto *et al*. (2011) criaram suas plataformas para monitorar e identificar ocorrências no trânsito baseado em dados coletados do *Twitter* a partir de dicionário de dados com termos referentes ao trânsito.

Recentemente o Waze preparou uma nova função para ajudar os turistas a evitarem rotas perigosas [TecMundo 2016]. O aplicativo retira informações da base de dados do Disque-Denúncia sobre áreas potencialmente perigosa. Assim, ao atravessar ou se aproximar de alguma dessas áreas, um ícone vermelho com um ponto de exclamação vai aparecer na tela

do aplicativo. O SigaCiente faz uma abordagem diferente, pois utiliza-se das mídias sociais para identificar rotas perigosas.

3. O SigaCiente

A proposta do trabalho é apresentar uma ferramenta para visualização das condições de trânsito e sugestão de rotas com observação de segurança. Para tanto, foram extraídos dados de mídias sociais *online* sobre as rotas de interesse. Também foram fornecidas funcionalidades para que os usuários do sistema colaborem fornecendo informações. Para descrever o sistema e seu funcionamento serão apresentadas as principais atividades e componentes do sistema proposto.

3.1. Arquitetura do Sistema

Os principais componentes do sistema estão demonstrados na Figura 1. O componente **Crawler** é responsável por obter informações sobre cidades, ruas, vias e proximidades em mídias sociais *online*. Esse componente depende das *APIs* externas fornecidas por essas mídias. A atividade de coleta de dados é constante e é realizada por regiões nas mídias sociais.

O componente **Definição de Rotas** é utilizada pelos usuários para eles informarem o trajeto de interesse. Os usuários podem informar textualmente a origem e o destino ou solicitar a rota a partir do local que estão utilizando o GPS do seu aparelho. Ele depende da interface *ObterRota*, pois ele fornece apenas os meios de interação com usuário. O componente **Inclusão de Informação** sobre rota permite que os usuários colaborem inserindo novas informações às rotas. Tais informações são salvas e serão utilizadas em próximas recomendações.

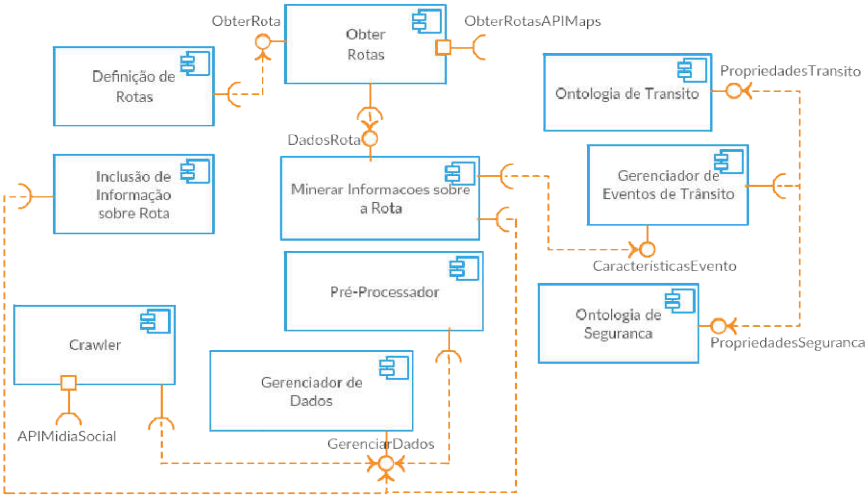


Figura 1. Diagrama de Componentes

O componente **Obter Rotas** depende de serviços de terceiros. Essa dependência está representada por *ObterRotasAPIMaps*. No presente trabalho, as definições de melhor rotas serão repassados a serviços de terceiros com *APIs* públicas disponíveis na Web. De posse das possíveis rotas recebidas (com informações de tempo de trajeto, trânsito e distância), o componente Obter Rotas levanta informações sobre a rota utilizando a base de dados do sistema. As informações levantadas buscam melhorar a experiência do usuário ao aproximar as informações da rota com o mapa cognitivo do mesmo. Neste componente também são

levantadas as informações de segurança sobre a rota (locais por onde passará o usuário). As informações de segurança serão utilizadas para reordenar as rotas recomendadas e para inserir informações sobre o grau de risco das mesmas.

O componente **Minerar Informações** é responsável por obter os dados pré-processados e extrair dados relacionados às rotas. A mineração busca por termos e características que são obtidas no componente **Gerenciador de Eventos de Trânsito** por meio da interface CaracteristicasEvento. Esse componente obtém informações sobre a rota utilizando **Ontologia de Trânsito** e **Ontologia de Segurança** que definem termos associados ao trânsito.

O componente **Pré-Processador** é responsável por remover conteúdo dos dados brutos (dados de comentários dos usuários ou de mídias sociais) que não interessam. Exemplos de pré-processamento são: remoção de *stop words, stemming*, remoção de caracteres especiais ou qualquer conteúdo que dificulte a mineração e/ou não agregue valor às informações sobre as rotas. Os dados pré-processados são utilizados para criação uma nova base de dados. O componente **Gerenciador de Dados** provê a interface *GerenciarDados* que possui funcionalidades de persistência e de recuperação dos dados.

Por utilizar funcionalidades de recuperação de dados utilizando *APIs* de mídias sociais, usar serviços de mapas *online* e acrescentar novas funcionalidades, o sistema proposto pode ser entendido como um *mashup* Web de fusão de dados para recomendação de rotas.

3.2. Método e Implementação do SigaCiente

As próximas seções descrevem o método utilizado para desenvolver o protótipo do sistema proposto observando a arquitetura apresentada. A recomendação e acréscimo de informações às rotas utilizam dados sociais de *crowdsourcing* e considera o grande volume de dados (*Big Data*) disponível nas mídias sociais conforme apresentado em [França *et al*. 2014].

Para implantar e avaliar o funcionamento do protótipo, foi definida uma área da cidade do Rio de Janeiro. A área definida abrange parte da Avenida Brasil, uma importante via da cidade. A delimitação é necessária devido a necessidade de se manter uma base de localidade e de correlacionar os dados sociais com essas localidades.

As seções que seguem descrevem a abordagem adotada em cada etapa da implementação da aplicação para que esta chegue, por sua vez, ao seu objetivo fim.

3.2.1. Coleta de Dados

Foram coletados dados do Twitter (*tweets*) usando um *crawler*. Para isso, utilizou-se a *Search API* que é retrospectiva e fornece resultados passados. As restrições impostas são: *tweets* são de até sete dias passados a partir da data de coleta; e máximo de 180 requisições a cada 15 minutos (resultando em no máximo 18 mil mensagens).

Foram definidos por observação no tema, uma região de interesse foi definida e a recuperação foi realizada com base na definição de um raio que cobrisse a área de interesse buscando não exceder tal área. Foram coletados aproximadamente 7 mil *tweets* em uma semana.

Os dados também foram obtidos por meio da colaboração social. A ferramenta também possui a funcionalidade para as pessoas interajam e informem incidentes sobre o trânsito e sobre criminalidade, contribuindo para a atualização e melhoria da ferramenta. A

extração de dados sobre o trânsito e ocorrências de crimes possibilitou que dados relevantes fossem expostos em um mapa.

3.2.2. Dicionário de Dados

Foi realizada uma pré-analise quando se percebeu a necessidade da construção de um processo de identificação semântica automatizado. Para essa análise foi coletados *tweets* de perfis de noticias sobre o trânsito do Rio de Janeiro: @bandnewsfmrj, @informerjo, @transitoriorj, @operacoesrio, @leisecarj, @gruponewsrj e @wazetrafficrio. O objetivo foi identificar conteúdo relevante automaticamente e independente da fonte. 1500 *tweets* foram extraídos para análise. Os *tweets* passaram por um pré-processamento simples para remoção de acentos e pontuação e então passaram por um algoritmo de contagem de palavras contidas no cenário do trânsito.

Para a construção do dicionário foi utilizada a ontologia TEDO [Redlich 2013] sobre eventos de trânsito. Toda a análise feita foi baseada nas classes e propriedades da ontologia. O dicionário de dados foi construído no formato de XML, onde todos os XML contêm termos e sinônimo sobre o trânsito e representa as seguintes classes da TEDO: Acidente, Enguiço, Eventos Climáticos, Engarrafamento, Interdição e Outros Eventos. Cabe ressaltar que nesta versão do protótipo não foi possível realizar a mesma análise para o cenário de crimes, pois nenhuma ontologia para esse domínio foi encontrada.

3.2.3. Pré-processamento

A execução do pré-processamento ocorre com o auxilio da biblioteca de recuperação de dados *Apache Lucene.NET*. Essa biblioteca oferece analisadores que permitem que sejam executadas técnicas de pré-processamento como tokenização, normalização, remoção de *stopwords* e *stemming.* Primeiro é utilizado um divisor que transforma todos os termos em *tokens,* então os normaliza e remove a pontuação. Em seguida, são removidos todos os acentos e *stopwords*. E por último é aplicada uma customização da biblioteca que permite reduzir as palavras flexionadas a sua raiz, retirando plural, sufixos, prefixos etc.

3.2.4. Identificação Semântica

Após o pré-processamento, a identificação semântica é a etapa demonstrada na Figura 2. Um analisador customizado da biblioteca *Lucene.Net* foi desenvolvido para permitir a varredura de XMLs, a indexação dos termos em memória e a comparação com o texto publicado na web.

Para cada termo extraído de um *tweet,* varre-se o XML na tentativa de identificar o local mencionado no texto e suas respectivas coordenadas. Prosseguindo a análise, a aplicação pega novamente cada termo do *tweet* para compará-los aos termos do XML que representa Eventos de Trânsito, com o propósito de identificar qual(is) evento(s) são noticiados naquela publicação. Por fim, são identificados fonte, dia e horário, que são recolhidos do próprio *tweet,* e então a direção do incidente e atores envolvidos. Ao terminar a análise, esses dados são salvos na base de dados pertinentes que o usuário terá acesso visualmente no mapa.

Utilizando o *Google Maps API,* um novo serviço foi criado para facilitar o acesso às ocorrências de trânsito que foram publicadas em diferentes fontes de dados, conforme visto na Figura 3. Dados sobre o trânsito coletados e filtrados, são exibidos de forma dinâmica e em tempo real em um mapa.

No cenário de 7000 *tweets* coletados em uma semana constatou-se que 500 *tweets* foram extraídos respeitando a ontologia. Dos *tweets* utilizáveis, 400 são de fontes de notícias (usuários mencionados anteriormente) que publicaram, foram mencionadas ou *retweetados.* Sendo assim, repetições de texto não foram eliminadas. Os outros 100 são de usuários comuns.

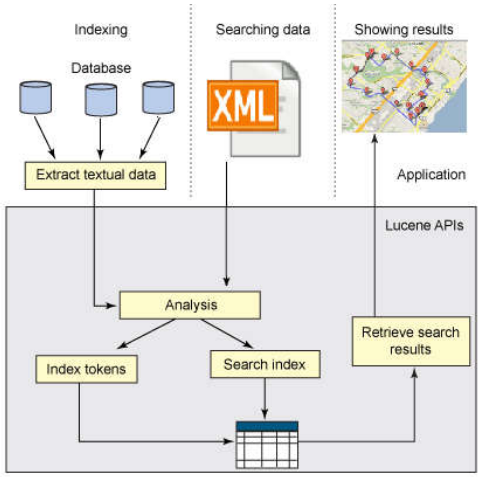


Figura 2. Funcionamento da Biblioteca *Lucene.Net*[3] (Adaptado de [Sonawane 2009])

Dos 500 *tweets,* obteve-se uma margem de acerto de 95%. Na margem de erro foram identificados *tweets* fora de contexto ou indicando locais errados, como por exemplo, um *tweet* que menciona a palavra padre e identifica um lugar chamado Padre Miguel.

3.2.5. Recomendação de Rotas

A recomendação de rota faz uso das rotas obtidas do *Google Maps.* Calcula-se o índice de criminalidade no percurso e sinaliza, através de variações de cores, as rotas mais e menos indicadas. Desde modo, o usuário ficará atento sobre os índices de criminalidade na região e poderá decidir qual caminho seguir, levando em consideração os riscos de prosseguir naquele trajeto. O cálculo utilizado será descrito de maneira funcional através do exemplo representado pela Tabela 1. A Figura 3 também ilustra a resolução do exemplo apresentado.

Os crimes são divididos em novos e antigos. Os novos são aqueles com menos de um ano. Calcula-se a probabilidade de um crime novo acontecer, levando em conta apenas os crimes de uma única rua. Serão chamadas de P(rua): Rua A – 0/5 = 0%; Rua B – 3 / 3 = 100 %; e Rua C – 1 / 2 = 50%. A probabilidade de um crime novo acontecer será calculada levando em conta apenas os crimes novos da região como neste exemplo P(novos): Rua A – 0/ 4 = 0%; Rua B – 3 / 4 = 75 %; e Rua C – 1 / 4 = 25%.

A probabilidade de um crime novo acontecer leva em conta apenas os crimes novos da região como um todo. P(região) é calculado da seguinte forma: Rua A – 0/10 = 0%; Rua B – 3 / 10 = 30 %; e Rua C – 2 / 10 = 20%. Então, calcula-se P(rua) – (P(novos) – P(região)). Seguindo o exemplo, isso seria calculado da seguinte maneira: Rua A – (0 - (0-0)) = 0%; Rua B – (100-(75-30)) = 55 %; e Rua C – (50-(25-20)) = 45%. E então, as cores simbolizam o

3 https://lucene.apache.org/core/

grau de risco: Vermelho – total >= 75%; Laranja – total >= 50% e total < que 75%; Amarelo – total >= 25% e total < que 50%; e Verde – total <25%.

Tabela 1. Exemplo de Crimes em uma Região

Ruas	Crimes Antigos	Crimes Novos	Total
A	5	0	5
B	0	3	3
C	1	1	2
Total	6	4	10

Cabe ressaltar, que a recomendação baseada em cores, ao invés de números, foi realizada para evitar a segregação de determinadas áreas. Tendo em vista que o cálculo é feito a partir das rotas indicadas pelo *Google Maps* e estas podem variar de acordo com imprevistos que retardam o trânsito, uma rota que é indicada como a menos recomendada em uma busca pode também ser a mais indicada em outra busca. Atualmente, o poder público já tem conhecimento dos riscos em determinadas regiões e a exposição poderia pressioná-los a tomar medidas que garantem a segurança da população. Como a proposta não é expor uma região como permanentemente perigosa, crimes com mais de um ano são considerados crimes antigos e tem impacto menor no cálculo.

Ainda sobre questão discriminação de ambientes, os governos publicam seus dados sobre crimes, como faz o governo do Estado do Rio de Janeiro[4]. Considera-se então que, do ponto de vista do impacto social (especulação no preço dos imóveis, discriminação, etc.), a proposta deste trabalho pode ter um potencial positivo ao permitir que a informação seja entendida facilmente pela população e os governos se sintam pressionados a observar as questões de segurança pública, apesar deste não ser o foco do presente trabalho.

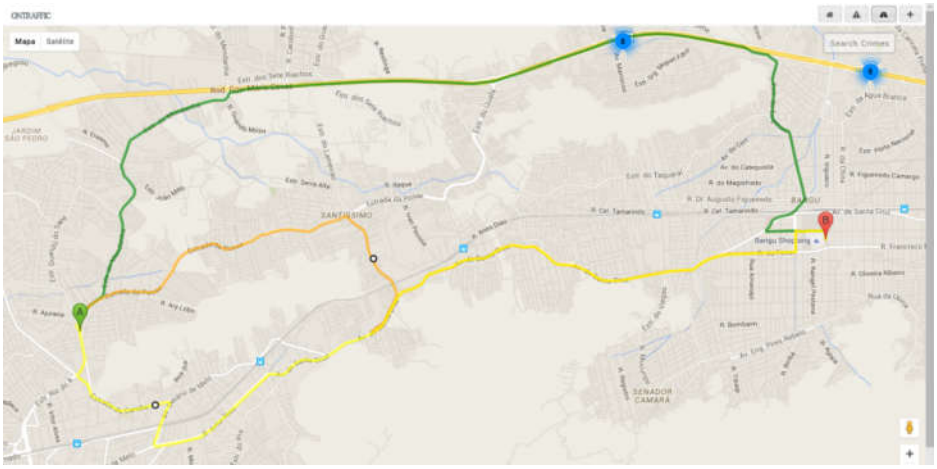


Figura 3. Recomendação de Rotas por Índice de Criminalidade

4. Considerações Finais

Este trabalho descreve o SigaCiente, um aplicativo que utiliza dados sociais para recomendar rotas considerando informações de segurança e para agregar novas informações às rotas. Percebeu-se, frente à literatura e o contexto de insegurança nas grandes cidades, especialmente no Brasil, a necessidade de considerar a segurança no trecho que será percorrido por um usuário de sistemas de recomendação de rota.

Considerando o levantamento na literatura, tornou-se perceptível o quanto essas novas informações são importantes para aproximar as rotas recomendadas dos mapas cognitivos dos usuários e de informações considerando o conceito de *wayfinding*. O protótipo desenvolvido nos permitiu trabalhar iterativamente melhorando o sistema proposto por possibilitar novos *insights* no tema. Cabe ressaltar, que o objetivo do trabalho foi agregar valor e inovação a ferramentas já existentes.

Como trabalho futuro pretende-se: avaliar a ferramenta que implementa a proposta do ponto de vista do usuário (testes da usabilidade, avaliação da interface, avaliação da experiência do usuário, entre outras); desenvolver um modelo ontológico para segurança do ponto de vista de criminalidade; bem como utilizar outras fontes de dados sobre crime (como dados do governo estadual e outras ferramentas web voltados para essa temática). Além disso, deseja-se avaliar e melhorar o mecanismo proposto de verificação de segurança da rota.

Referências

Bastos, A. V. B. Mapas cognitivos e a pesquisa organizacional: explorando aspectos metodológicos (2002). Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-294X2002000300008>. Acessado em 1 de Setembro de 2016.

Darlington, S. (2015) "Waze app directions take woman to wrong Brazil address, where she is killed", In: CNN, http://edition.cnn.com/2015/10/05/americas/brazil-wrong-directions-death, Outubro.

Endarnoto, S. K. *et al.* (2011) "Traffic Condition Information Extraction & Visualization from Social Media Twitter for Android Mobile Application", In: International Conference on Electrical Engineering and Informatics.

Franco, L. S. C. (2013) "Twittraffic: Uma plataforma de monitoração, visualização e identificação de ocorrências no trânsito". Dissertação de Mestrado. UFOP.

Lima, V. G. *et al.* (2012) "UbibusRoute: Um Sistem de Identificação e Sugestão de Rotas de Ônibus Baseado em Informações de Redes Sociais", In: VIII Simpósio Brasileiro de Sistemas de Informação.

Redlich, L. R. (2013) "Modelagem de eventos de trânsito com base em clipping de grandes massas de dados da Web". Dissertação de Mestrado. PUC/RJ.

Silva, J. B. *et al.* (2015) "*Wayfinding* em Aplicativos de Recomendação de Rota: Coerência com Mapas Cognitivos", In: Anais do 15º Ergodesign & Usihc.

Sonawane , A. (2009) Usando o Apache Lucene para Procura de Texto Senior. Disponível em <https://www.ibm.com/developerworks/br/java/library/os-apache-lucenesearch/>. Acessado em 10 de Maio de 2016.

TecMundo. (2016) "Nova função do Waze evita áreas de com maior risco de crimes no Rio", In: TECMUNDO, http://www.tecmundo.com.br/apps/107940-nova-funcao-waze-evita-areas-maior-risco-crime-rio.htm, Outubro.

Avaliando uma Estratégia Computacional Baseada em Workflows Científicos Apoiados por Placas Gráficas Genéricas

Fábio da Silva Cardozo[1], Ulisses Roque Tomaz[1], Sergio Manuel Serra da Cruz[1,2]

[1] Universidade Federal Rural do Rio de Janeiro (PPGMMC/UFRRJ)
[2] Programa de Educação Tutorial (PET-SI/UFRRJ)
BR-465, Km 7 Seropédica-Rio de Janeiro-RJ-Brazil

{ulisses.rtomaz, fcardozo}@gmail.com, serra@pet-si.ufrrj.br

Resumo. *O crescente volume de dados necessários para a realização de pesquisas atmosféricas e climáticas oferecem desafios. Este trabalho tem como objetivo apresentar uma estratégia computacional baseada em workflows científicos e técnicas de consistência dados destinada ao tratamento de longas séries de dados pluviométricos, para isso utiliza-se recursos de computação paralela em placas gráficas de propósito geral com coleta de dados de proveniência retrospectiva. Os primeiros resultados mostram que a estratégia apresenta altas taxas de preenchimento de falhas e crescentes ganhos de desempenho quando comparada a abordagem sequencial tradicional.*

Abstract: *The increasing volume of meteorological data needed to conduct atmospheric and climate studies offer new challenges for data analysis. This work aims to present the initial steps of a computational approach. It is tailored to compute long series of rainfall data using parallel computing capabilities of low cost and general purpose GPU graphics cards. The approach uses retrospective provenance to enrich the quality of the raw rainfall data. Our initial results show that the approach not only augment the quality of the data but also offer performance gains when compared to the traditional sequential approach.*

1. Introdução

As ciências da terra têm avançado velozmente graças a experimentação computacional, tendo como um ponto importante os estudos baseados em simulações numéricas dos fenômenos meteorológicos (BATCHELOR, 2000). As mudanças climáticas em curso podem resultar em impactos agroambientais severos. Aumentos nas temperaturas do ar são esperados em nível mundial e uma das consequências decorrentes serão as variações nos ciclos hidrológicos (IPCC, 2013). Acredita-se que a maior ameaça para os seres humanos será manifestada em nível local, através de mudanças em eventos regionais de tempo e clima extremos. O Brasil é particularmente vulnerável a mudanças na frequência e intensidade de eventos extremos, como ondas de calor, secas, enchentes e chuvas extremas, como ocorrido nos últimos anos nas regiões Sul e Sudeste.

Estudos meteorológicos são caracterizados por manipularem grandes quantidades de dados em longas séries contínuas e consistentes. Contudo, obter séries com essas características ainda é um grande desafio nesta área de estudo. As falhas e perdas dos dados ocorrem desde o momento da coleta na estação meteorológica até sua disponibilidade em repositórios de dados (LEMOS FILHO *et al*., 2013). Além dessas dificuldades, os dados se encontram em diferentes estruturas, formatos, com descontinuidades cronológicas e sem os descritores de proveniência.

Esse trabalho tem como objetivo apresentar uma estratégia computacional centrada no uso técnicas de consistência de dados pluviométricos apoiadas por *workflows* científicos executados em ambientes paralelos que utilizam placas gráficas do que são capazes de transformar dados pluviométricos brutos em dados curados de qualidade enriquecidos por proveniência (FREIRE et al. 2008). A estratégia apresentada é capaz de processar tais dados em arquiteturas *Compute Unified Device Architecture* (CUDA), que envolvem computação paralela em placas de processamento gráfico (GPGPU) de baixo custo e de uso genérico.

2. Trabalhos Relacionados

Atualmente, ambientes paralelos baseados em GPGPU representam uma alternativa para acelerar aplicações científicas baseadas em *workflows* científicos (GOSWAMI et al, 2016, LIU et al. 2016). Estes ambientes possibilitam que problemas complexos, de simulação computacional, sejam executados em tempos reduzidos e a baixos custos quando comparados com os computadores de alto desempenho. Para este fim, os sistemas heterogêneos compostos de processadores *multi-core* e aceleradores *many-core* como as GPGPUs representam uma tendência na atualidade. No entanto, o maior desafio está em identificar quais são as atividades do workflow ou os trechos de código são os mais adequados para cada tipo de arquitetura. Nesta seção apresentamos os fundamentos utilizados pela estratégia MetFlow concebida por (CARDOZO, 2014).

2.1 Técnicas de Consistência e Preenchimento de Falhas em Dados Pluviométricos

O controle de qualidade de dados meteorológicos adota sequencias de filtros de dados (FENG *et al.*, 2004), que são aplicados na detecção e identificação de erros nos dados brutos coletados pelos sensores das estações meteorológicas. Segundo Magina (2007) os filtros utilizados são capazes de identificar registros extremos reais e registros espúrios, os últimos devendo ser excluídos, pois se mantidos na série histórica aumentariam a frequência de casos extremos, e distorceriam a estimativa de parâmetros que produzem a função de probabilidade de extremos e os tempos de retorno.

Por adequação ao trabalho proposto as regras estabelecidas por Feng *et al.* (2004) foram adotadas nessa pesquisa, as quais foram aplicadas às séries de dados produzidas pelas estações meteorológicas na detecção de valores suspeitos. As regras utilizadas são: (i) Detecção de Valores Extremos Máximos e Mínimos; (ii) Verificação de Eventos Temporalmente Isolados; (iii) Verificação de Eventos Espacialmente Isolados.

Outra etapa do controle de qualidade dos dados meteorológicos consiste em realizar o preenchimento dos dados ausentes da série por meio de métodos estatísticos. Os métodos mais utilizados são: (i) Regressão linear simples (RL); (ii) Vizinho mais

próximo (VP); (iii) Ponderação Regional com Bases em Regressões Lineares (PRRL); (iv) Ponderação Regional (PR) e (v) Inverso da Potência da Distância (IPD). Precinoto *et al.* (2013), em estudos anteriores, verificaram que o método RL é um dos mais adequados para preenchimento de falhas sobre dados de precipitação na região Sudeste do estado do Rio de Janeiro. Portanto, este será o adotado neste trabalho.

2.2 Tecnologias GPGPU-CUDA

Nos últimos anos as placas gráficas GPGPU tiveram seu uso amplamente difundido. Verificou-se que seu poder computacional, incialmente disponível para jogos, era potencialmente aplicável para resolução de diversas categorias de problemas científicos (CHAKRABARTI *et al.*, 2012, GOSWAMI et al., 2016).

Adotou-se para este trabalho a arquitetura CUDA da NVIDIA. Este ambiente de programação que permite realizar computação de propósito geral utilizando a GPU e fornece acesso aos recursos do hardware através de comandos semelhantes aos das linguagens C. Os módulos CUDA implementam *threads* paralelas para executar as atividades do *workflow* relacionadas com as análises de dados e preenchimento de falhas baseados no método RL de preenchimento de falhas.

2.5. MetFlow

O MetFlow, cuja primeira versão foi concebida por Cardozo (2014), é um *workflow* científico que realiza tratamento de dados e prognósticos quantitativos sobre dados de precipitação pluvial. Tais prognósticos utilizam grandes volumes de dados representados por longas séries temporais.

Por intermédio do MetFlow se executam experimentos do tipo simulações numéricas dos fenômenos meteorológicos distribuídos em ambientes de computação sequenciais e paralelos e também se armazenam dados curados juntamente com seus metadados de proveniência retrospectiva de cada execução.

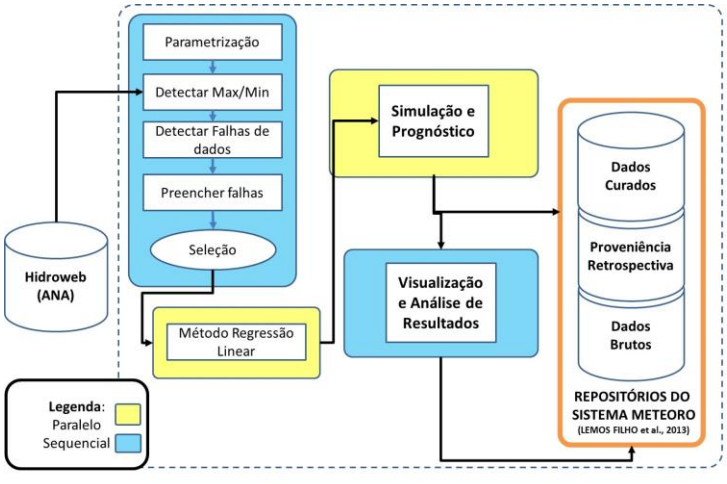


Figura 1 – Representação conceitual das fases de experimentos computacional em Pluviometria apoiada pelo MetFlow nos ambientes sequencial e paralelo.

A atual versão do MetFlow utiliza módulos CUDA e python, apoiada pelo método de regressão linear (RL) para efetuar preenchimento de falhas das séries temporais

2.6 Trabalhos Relacionados

Atualmente, existem diversos trabalhos na área de pré-processamento de dados meteorológicos. Magina (2007) propõe de forma semiautomática formatos de tratamento estatísticos para series históricas meteorológicas, mas conta somente com o suporte de macros em planilhas MS Excel para aplicação desse tratamento. Esta abordagem torna o trabalho humano massivo e sujeito a erros. Além disso, não incorpora as questões de proveniência de dados. Lemos Filho *et al.* (2013), propõem um sistema baseados em proveniência e *pipelines* de pré-processamento de dados pluviométricos em uma plataforma Web, que, no entanto, não traz uma solução ou aplicação paralela de alto desempenho, nem se utiliza de *workflows* científicos. A Tabela 1 apresenta uma comparação entre as funcionalidades de trabalhos correlatos.

Asvija *et al.* (2010) sugerem o uso de *workflows* científicos para implementar modelo numérico meteorológico *fifth-generation Model Mesoscale*, que trata do prognóstico em Meteorologia de mesoescala de fenômenos atmosféricos, como brisas, tempestades de convecção, não contemplando o tratamento de dados e coleta de dados de proveniência.

Horta *et al.* (2013) trata do uso de *workflows* científicos empregados em ambiente de clusteres, mas não apresentam solução desenvolvida que se aplique ao problema que norteia este trabalho. Os autores se restringem ao campo da investigação teórica, indicando o que é possível realizar juntando *workflow* científico e computação massiva paralela com GPU.

Tabela 1 – Comparativo entre os trabalhos relacionados.

	MetFlow (2014)	Filho et al. (2013)	Magina (2007)	Horta, et al. (2013)	Asvija et al. (2010)	Whang e Shi (2014)
Uso de Workflows científicos	X	-	-	X	X	-
Coleta de proveniência retrospectiva	X	X	-	-	-	-
Técnicas de detecção de falhas e preenchimento	X	X	X	-	-	X
Uso de técnicas de paralelismo de dados	X	-	-	-	X	X
Uso de GPGPU	X	-	-	X	-	X
Programação em CUDA	X	-	-	X	-	X
Prevê validação cruzada de dados	X	-	-	-	-	-
Uso de dados de pluviometria	X	X	X	-	-	-
Compartilhamento de esquemas relacionais	X	X	-	-	-	-
Compatibilidade com *schema* PROV	X	-	-	-	-	-

Whang e Shi (2014) reconhecem o problema da qualidade de dados meteorológicos e apresentam uma metodologia e um *workflow* (sem uso de SGWfC) baseado em MPI executado em placas GPGPU para tratar dados sumarizados de

temperatura da superfície da Terra coletados pela WMO/NOAA, a iniciativa gera dados curados em formato texto estruturado facilitando simulações computacionais. No entanto, o trabalho não considera a importância da proveniência de dados.

3. Metodologia

Os dados utilizados nesse trabalho são oriundos do sistema HidroWeb mantido pela Agência Nacional de Águas (HIDROWEB, 2016). Eles são séries temporais pluviométricas não consistentes sobre: (i) históricos diários de chuva; (ii) inventário de bacias, rios, estados, municípios, estações (pluviométricas e fluviométricas) e suas respectivas coordenadas geográficas.

Para a execução dos experimentos foi utilizado um notebook com processador Intel Core i5 dual core de 2,9 GHz, com 4GB de RAM DDR3 de 533 MHz e placa de vídeo (GPGPU) NVIDIA GeForce GT 335M de 1080 MHz, com 1GB e 72 cores CUDA. Neste estudo utilizamos dados de chuva da região Sul Fluminense, pois apresentam importância industrial, agropecuária e também por ser a principal fonte de captação de águas do estado do Rio de Janeiro. Dentre todas as estações da região, foram selecionadas todas as 77 estações meteorológicas dentro da área de maior pluviosidade e de altimetria próximas. As coordenadas dessas estações variam entre as latitudes são 22º 03’ e 23º 21’ S e longitudes 43º 25’ e 44º 54’ W.

Consideraram-se apenas séries de chuvas superiores ou iguais a 20 anos de dados e com início a partir de janeiro de 1960 até dezembro de 2013. Ou seja, nossos experimentos utilizaram séries de dados pluviométricos reais com 53 anos. Como prova de conceito utilizou-se o workflow MetFlow composto por um conjunto de atividades (módulos) que fazem parte do *workflow* científico paralelo que é capaz de processar longas séries de dados pluviométricos.

De acordo com a Figura 1, os experimentos realizados por intermédio do MetFlow podem divididos em três fases distintas: 1) *pré-processamento, 2) prognóstico e 3) visualização* dos dados meteorológicos. Os dados brutos processados pelo workflow e os dados curados são armazenados em um repositório de dados do tipo relacional capaz de armazenar os metadados de proveniência retrospectiva gerados a cada execução do mesmo.

Este tipo de abordagem é muito importante, pois permite consultar e compartilhar conjuntamente os dados e proveniência retrospectiva dos experimentos, ampliando sua transparência e confiabilidade. Ressalta-se que o repositório de dados utilizado na pesquisa compartilha o mesmo esquema originalmente proposto por Lemos Filho *et al.* (2013).

4. Experimentos em Sequencias e Paralelos com Regressão Linear

Inicialmente, os experimentos carregam os dados pluviométricos brutos do HidroWeb através do MetFlow, que é composto por um conjunto de tarefas configuradas pelo pesquisador [parametrização, preenchimento de falhas, preparação de dados para o ambiente paralelo (vetorização) e processamento de dados] executados em ambientes sequencial e paralelo. O modelo estatístico RL efetua o preenchimento de falhas das séries pluviométricas é processado em paralelo por meio de código CUDA, gerando

novos repositórios de dados enriquecidos por metadados de proveniência que etiquetam cada item de dado analisado resultam desse processo.

Em especial, a proveniência (FREIRE et al. 2008) atua como um certificado de qualidade e autenticidade de cada item de dado meteorológico, o que permite o compartilhamento e reuso dos dados sem falhas com descritores detalhados, que assim, explicitam a lógica da geração de cada item de dado do banco.

Neste trabalho foram utilizadas duas versões do *workflow* MetFlow no VisTrails. A versão sequencial possui módulos de processamento desenvolvidos em python e a versão paralela possui módulos adicionais de processamento paralelo desenvolvidos em linguagens CUDA e Python. Ambas utilizaram o MySQL acoplado ao VisTrails. Os dados, assim como, a proveniência retrospectiva são armazenados na base de dados do sistema Meteoro (LEMOS FILHO *et al.*, 2013) e podem ser armazenados em granularidades distintas.

Os módulos sequenciais são codificados apenas em python e os paralelos em linguagem CUDA (seus códigos estão disponíveis mediante solicitação aos autores). Os módulos paralelos são aqueles que consomem mais recursos de processamento, ou seja, vetorização e preparação da regressão linear por esse motivo são executados em ambientes de maior capacidade de processamento. O módulo coletor de proveniência utilizado no protótipo foi o disponível internamente pelo VisTrails (CALAHAN et al., 2006).

Uma das principais características do MetFlow são os elevados percentuais de correção e preenchimentos de falhas e baixo tempo de execução quando comparados com a tradicional abordagem manual. Em termos de quantificação de correções de falhas, os resultados estão apresentados na Tabela 2, onde é possível observar os números de meses com falhas e os percentuais de ajustes dos dados. Por exemplo, nos experimentos com 36 estações (dados desde 1960 até 2013, ou seja, 53 anos de dados e um total de 22.896 meses) haviam 1.461 meses que apresentavam algum tipo de falhas, sendo corrigidos 1.310 meses.

O MetFlow em sua versão paralela utilizando apenas três *threads* e o método RL com 36 estações dentro de um raio de 20Km de distância foi capaz de corrigir automaticamente 89,6% dos casos (1.310 meses) com um tempo médio de processamento de 5:57m, enquanto que o percentual de acerto para 77 estações foi de 72,17% com um tempo médio de processamento de 16:35m. Os valores percentuais crescentes de correções se devem ao maior número de falhas corrigidas nas séries de dados e ao maior número de estações. As diferenças de tempos de execução entre as versões sequencial e paralela são pequenas, porém crescentes e ligeiramente melhores na versão paralela.

Tabela 2 – Variações percentuais de correções de falhas executadas pelo MetFlow em sua versão paralela usando o método regressão com raio de distância de 20Km e número de execuções de cada versão do MetFlow = 5.

	Meses avaliados (= 12 * nº. estações * 53 anos)	Meses com falhas	Meses com correções realizadas	Tempo médio de execução sequencial	Tempo médio de execução paralelo	% de acertos dos dados
17 estações	10.812	552	327	2:53m	2:51m	59,24
36 estações	22.896	1.461	1.310	6:10m	5:57m	89,66
77 estações	48.972	2.479	1.789	17:51m	16:35m	72,17

5. Conclusão

Neste trabalho ficou clara a adequação da modelagem de um experimento em Pluviometria para o tratamento e preenchimento de falhas usando o MetFlow em ambiente baseado em GPGPU. Porém, nossos primeiros resultados mostram que para atingir uma solução paralela mais genérica se necessitam de investigações complementares.

As pesquisas com workflows científicos em execução paralela com coleta de proveniência em placas genéricas do tipo GPGPU ainda estão se desenvolvendo no Brasil e no mundo, ainda não há um referencial teórico estabelecido. No entanto, estimamos que seu o uso de placas gráficas para acelerar a execução de workflows será amplamente desenvolvido futuramente.

As principais contribuições deste trabalho são o expressivo percentual de correção e preenchimento de falhas pelo MetFlow aliados com geração de anotações de proveniência enriquecendo os dados. No que tange à correção das falhas obtivemos acertos expressivos com valores superiores a 59 % mesmo com pequeno número de estações, este valor pode se ampliar à medida que se aumentam o número de estações e se variam as distâncias entre eles, que são estabelecidos nos parâmetros utilizados pelos cientistas ao executar o MetFlow.

Dentre as principais limitações destacamos que os ganhos de desempenho da versão paralela comparada com a versão sequencial foram pouco significativos, indicando que maiores estudos devem ser realizados para verificar a adequação do uso ou da escolha do módulo paralelizável ou mesmo a escolha de melhores equipamentos para a realização dos experimentos.

Agradecimentos

Agradecemos ao FNDE e ao MEC/SeSU pelo financiamento do programa PET-SI/UFRRJ e ao programa PPGMMC/UFRRJ pelas bolsas concedidas.

Referências

Asvija, B.; Shamjith, K.V.; Sridharan, R.; Chattopadhyay, S. Provisioning the MM5 meteorological model as Grid Scientific Workflow. 2010.

Batchelor, G. K. An Introduction to Fluid Dynamics. Cambridge University Press. 2000.

Cardozo, F. S. Tratamento e preenchimento de falhas de séries de dados meteorológicos utilizando *workflows* científicos paralelos em ambientes de GPU. Dissertação Mestrado – UFRRJ, 2014.

Callahan, S. P.; Freire, J.; Santos, E.; Scheidegger, C. E.; Silva, C. T.; Vo, H. T. VisTrails: visualization meets data management. In: Proceedings of the, 2006ACM SIGMOD, p. 745-747, 2006.

Chakrabarti, G.; Grover, V.; Aarts, B. et al. CUDA: Compiling and optimizing for a GPU platform. Procedia Computer Science, v. 9, p. 1910–1919. 2012.

Deelman, E.; Gannon, D.; Shields, M.; Taylor, I. Workflows and e-Science: An overview of workflow system features and capabilities, FGCS, v. 25, n. 5, p. 528-540, 2009.

Freire, J.; Koop, D.; Santos, E.; Silva, C.L. Provenance for computational tasks: A Survey. Computing in Science & Engineering, v. 10, n. 3, p. 11–21, 2008.

Feng, S.; Hu, Q.; Qian, W. Quality control of daily meteorological data in China, 1951-2000: a new dataset. Int. Journal of Climatology, v. 24, n. 7, p. 853–870. 2004.

Goswami, A. et al., Landrush: Rethinking In-Situ Analysis for GPGPU Workflows, 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), Cartagena, pp. 32-41, 2016.

Lemos Filho, G. R.; Precinoto, R. S.; Correia, T. P.; Santos, E. O.; Lyra, G. B.; Cruz, S. M. S., Assimilação, Controle de Qualidade e Análise de Dados de Meteorológicos Apoiados por Proveniência, VII e-science Workshop, XXXIII CSBC. 2013;

Liu, J. et al. An efficient geosciences workflow on multi-core processors and GPUs: a case study for aerosol optical depth retrieval from MODIS satellite data. 2016. http://dx.doi.org/10.1080/17538947.2015.1130087.

Hidroweb Sistema HidroWeb. 2015. Disponível em: <http://hidroweb.ana.gov.br>. Acesso em 14 de maio de 2014.

Horta, F.; Dias, J.; Elias, R. et al. Prov-Vis: Large-Scale Scientific Data Visualization Using Provenance. 2013.

IPCC Climate Change 2013: The Physical Science Basis. Disponível em: <http://www.ipcc.ch/report/ar5/wg1/>. Acesso em 14 de março de 2016.

Magina, F. C. Aquisição Automática e Tratamento de Dados Meteorológicos Aplicáveis ao Projeto e Operação de Linhas Aéreas de Transmissão de Energia Elétrica. Dissertação de Mestrado. 2007.

Precinoto, R. S.; Lemos Filho. G. R.; Correia, T. P.; Santos, E. O.; Lyra. G. B.; Cruz, S. M. S. 2013. Uso De Sistema De Pré-Processadores Para Obtenção De Séries Pluviométricas De Qualidade. Congresso Brasileiro de Agrometeorologia 2013.

Shi, X.; Wang, D. Processing NOAA Observation Data over Hybrid Computer Systems for Comparative Climate Change Analysis. In: WorldComp, 2014. p. 1.

Busca semântica aplicada à recuperação de informações de contexto histórico

Geovani S. Celebrim[1], Ricardo L. S. Melo[1], Alexandre Fortes[1], Leandro G. M. Alvim[1], Luis F. Orleans[1]

[1] Departamento de Ciência da Computação – Instituto Multidisciplinar
Universidade Federal Rural do Rio de Janeiro (UFRRJ)
Nova Iguaçu – RJ – Brasil

geovanicelebrim@ufrrj.br, ricardoluiz@ufrrj.br,

fortes.ufrrj@gmail.com, alvim.lgm@gmail.com,

lforleans@ufrrj.br

Abstract. *Before the web, it was a challenge to historians to find historical sources for the research work. In recent decades, the problem was reversed. It became impractical to read all available sources and, thus, the data mining became fundamental in supporting the search in the domain. Here, we propose to the history domain, a semantic search engine, which considers the historical context and meaning of the queries. We show, for instance, that it is possible to identify relations between people and a particular historical event, or even statistics on a particular event. In this work, we indicate that the semantic search adapted to the domain is able to transform the research in the field of History.*

Resumo. *Antes da web, era um desafio para os historiadores encontrar fontes históricas para o trabalho de pesquisa. Nas últimas décadas, o problema se inverteu. Tornou-se inviável ler todas as fontes disponíveis e, assim, a mineração de dados tornou-se fundamental ao apoio na pesquisa do domínio. Aqui, propomos para o domínio da história, um buscador semântico que considera o contexto histórico e o significado das consultas. Mostramos, por exemplo, que é possível identificar relações entre pessoas e um determinado evento histórico, ou até mesmo estatísticas sobre um determinado acontecimento. Nesse trabalho, indicamos que a busca semântica adaptada ao domínio é capaz de transformar a pesquisa no campo da História.*

1. Introdução

Os buscadores semânticos surgiram visando melhorar a precisão das buscas convencionais, trazendo um conjunto de informações relevantes baseado no contexto e domínio de busca [Guha and McCool 2003]. Antes do seu surgimento, a qualidade dos resultados dependia, na maioria das vezes, de como o usuário elaborasse a busca. Após seu surgimento, os resultados de buscas passaram a ser mais precisos pois o contexto semântico passou a ser considerado. O principal benefício da utilização de buscadores semânticos para a História dá-se pela tentativa de se descobrir a real intenção do usuário e não apenas se baseando em métodos de similaridade de termos, como é feito nos buscadores convencionais. Os resultados dos buscadores semânticos além de serem mais precisos, trazem

também informações adicionais. Tais características são fundamentais para a garantia de qualidade dos resultados de pesquisas relacionadas ao domínio da História.

O repositório textual do Centro de Documentação e Imagem [CEDIM 2016] (CEDIM) busca tornar-se um espaço de pesquisa tanto para o público acadêmico quanto o público em geral. O CEDIM caracteriza-se por um canal sistematizado para a reunião e disponibilização de documentação visual, iconográfica e sonora. Seu acervo é composto basicamente por documentos digitalizados e entrevistas já realizadas por pesquisadores. Buscando elaborar um mecanismo eficiente para a recuperação de informações, propomos neste trabalho, um buscador semântico para o domínio da História.

O presente trabalho está estruturado em seis seções principais. Na seção 2, um referencial teórico sobre busca semântica, suas metodologias e reconhecimento de entidades é apresentado. A seção 3 apresenta abordagens utilizadas por diferentes buscadores semânticos na literatura. Na seção 4, um estudo de caso sobre a metodologia de pesquisa de documentos no repositório CEDIM é apresentado. Na seção 5, a arquitetura do buscador semântico proposto para o repositório CEDIM é detalhada. A seção 6 apresenta alguns resultados a fim de mostrar os benefícios e vantagens da busca semântica. Por fim, na seção 7, o trabalho é sumarizando apresentando conclusões e trabalhos futuros.

2. Busca semântica

Diferentemente dos motores de busca convencionais que utilizam algoritmos de similaridade de termos e ranqueamento de páginas, os buscadores semânticos baseiam-se no significado contextual das palavras em um domínio semântico ou um modelo de conhecimento [Renteria-Agualimpia et al. 2010]. Adicionalmente, os buscadores semânticos visam melhorar os resultados das buscas, reunindo um conjunto de informações relevantes e que estejam de acordo com os objetivos do usuário.

Segundo [Guha et al. 2003] existem dois tipos de buscas: as de navegação e as de pesquisa. Na primeira, o objetivo é localizar uma página ou documento específico. Já na segunda, procura-se encontrar um conjunto de páginas e documentos que juntos fornecem amplo conhecimento sobre o assunto pesquisado. Analisando os dois tipos, pode se afirmar que a busca semântica se caracteriza por uma busca de pesquisa, já que ao invés de buscar um dado ou página específica, ela procura agrupar dados relevantes que trazem conhecimento sobre o que foi buscado.

2.1. Abordagens

Na literatura, algumas abordagens sobre buscadores semânticos podem ser encontradas [Mäkelä 2005]. Os buscadores voltados para a *web* em sua grande maioria se baseiam no modelo *Resource Description Framework* (RDF). O RDF é um padrão para a troca de recursos na *web* que visa a interoperabilidade semântica entres diferentes sistemas digitais [W3C 2014]. Utilizando meta-dados para descrever um recurso ou relações entre recursos, o RDF permite criar um grafo a partir dos dados, representando recursos e suas relações como nós e arestas. Além dos buscadores baseados na *web*, exitem aqueles fundamentados em repositórios de dados que, geralmente, são restritos a um domínio.

Segundo [Ramachandran and Sujatha 2011], dentre as diversas abordagens utilizadas como base para buscas semânticas, as seguintes se destacam: (i) RDF *Path Traversal*, busca informações relevantes adicionais a partir de um nó inicial de um grafo

ponderado formado pelo modelo de dados RDF, no qual os pesos das arestas refletem a importância dos relacionamentos; (ii) *Keyword-Concept Mapping*, cria um mapa conceitual a partir da detecção de palavras-chaves de buscas em linguagem natural; (iii) *Graph patterns*, metodologia utilizada com o intuito de criar conexões relevantes entre os dados, sendo mais utilizado em visualização de dados; (iv) *Logics*, aplica regras de inferência utilizando como base a *Web Ontology Language* (OWL), uma linguagem ontológica para a *web*. Apesar da OWL ser baseada em lógica descritiva, essa metodologia dificilmente é utilizada devido a necessidade de se processar grandes volumes de dados.

2.2. Reconhecimento de Entidades Nomeadas

O Reconhecimento de Entidades Nomeadas (REN) pode ser definido como o processo de extração e classificação de informações de regiões de um texto, que corresponde ao nome de uma entidade [Marrero et al. 2013]. Portanto, são ditos entidades nomeadas expressões que nomeiam locais, quantidades, pessoas e organizações [Zhou and Su 2002]. A tabela 1 apresenta alguns exemplos de entidades nomeadas que podem ser extraídas de um determinado documento.

Nomenclatura	Descrição
Pessoa	Nomes próprios de pessoas, sendo estes nomes completos ou parciais. (ex.: Julio do Valle; Julio; Valle)
Evento	Citação de eventos históricos importantes. (ex.: Segunda Guerra Mundial).
Local	Nomes próprios de locais (ex.: Rio de Janeiro; Alemanha).
Data	Datas, completas ou parciais, de acontecimentos históricos (ex.: 11 de novembro de 1942; maio de 1945).
AutorReporter	Nomes próprios de autores e/ou repórteres envolvidos na elaboração do noticiado.
Pesquisador	Nomes próprios de pesquisadores citado no noticiário.
Organização	Nomes de organizações, empresas, instituições, etc. (ex.: FEB; Polícia Militar).
Fonte	Citação da fonte da informação.
TempoFonte	Data da publicação da notícia.
URLFonte	Endereço da fonte disponibilizado pela rede.
Artefato	Nome dado a um mecanismo construído para um fim determinado. (ex.: Bomba Atômica)
Quantidade	Indicador de quantidade para algum evento pertinente (ex.: 8 mil mortos; mais de 100 pessoas feridas).

Tabela 1. Entidades nomeadas.

Existem diversas técnicas de extração automática de entidades nomeadas, uma muito utilizada é a chamada *Conditional Random Fields*, que trata-se de um modelo probabilístico que busca rotular e segmentar os dados [Lafferty et al. 2001]. As técnicas de reconhecimento automático de entidades requerem a construção de um corpus de treinamento. Para a construção desse corpus, as entidades são anotadas com o apoio de ferramentas como o BRAT [Stenetorp et al. 2012].

A figura 1 apresenta um exemplo de anotação de uma entidade nomeada realizada pela ferramenta BRAT. Ao final do processo de anotação, é gerado um arquivo que possui

a classe que a entidade anotada pertence, o intervalo de caracteres ocupado pelo trecho anotado e o próprio trecho anotado. Neste arquivo, encontram-se também as possíveis relações entre entidades anotadas, como pode ser visto na figura 2.

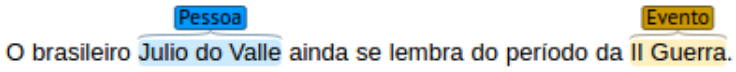


Figura 1. Exemplo de entidade nomeada.

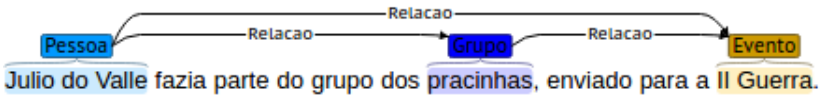


Figura 2. Exemplo relação entre entidades nomeadas.

3. Trabalhos Relacionados

Enquanto a maioria do buscadores semânticos encontrados na literatura focam em buscas de páginas *web*, existe um subconjunto que foca em repositórios de dados pertencentes a um domínio específico ou em formatos de textos estruturados. Nesta seção, as abordagens referentes a buscadores semânticos pertencentes a esse subconjunto estão descritas nos parágrafos subsequentes.

Kleio é um buscador semântico feito especialmente para um repositório de documentos no domínio da medicina [Nobata et al. 2009]. Ele utiliza metadados para indexar documentos por conceitos semânticos referentes a termos biomédicos como genes, proteínas, dentre outros termos relacionados. O reconhecimento de termos é realizado por meio da técnica de reconhecimento de entidades, e então os índices são gerados por meio de ferramentas de indexação de termos.

Bibster é um buscador semântico feito para o *DBLP Computer Science Bibliography* (DBLP), repositório *on-line* que indexa informações bibliográficas sobre publicações da área de ciências da computação [Haase et al. 2004]. Ele utiliza os dados já indexados disponibilizados no DBLP como base para a criação do buscador semântico. Além disso, utiliza uma rede *peer-to-peer* a fim de reduzir o número de acessos ao DBLP.

XSearch é um buscador semântico para documentos XML que, diferentemente dos buscadores convencionais, não retorna arquivos completos, mas sim, fragmentos [Cohen et al. 2003]. A vantagem de apenas retornar fragmentos deve-se ao fato de que um documento completo na maioria das vezes contém informações adicionais irrelevantes ao contexto de busca. O *XSearch* utiliza *tags* de arquivos XML para realizar o processo de indexação, e faz uso de técnicas de *information-retrieval* a fim de ranquear os resultados relacionados semanticamente aos termos de busca.

O *History Lab* é um portal que disponibiliza diversas ferramentas para a realização de buscas semânticas no domínio da história [History-Lab 2016]. As buscas são realizadas sobre uma base de documentos históricos, muitos deles digitalizados, pertencentes a departamentos do governo estadunidense. As ferramentas utilizam técnicas de *data mining* e *information retrieval* com o intuito de proporcionar meios de se realizar buscas contextuais baseados em tópicos, pessoas, países e datas. Além disso, realiza buscas por palavras-chave e disponibiliza diversos meios para a visualização dos resultados.

4. Estudo de Caso

Este trabalho surgiu a partir da necessidade de ferramentas que apoiem os historiadores do CEDIM durante os procedimentos de pesquisa em seu vasto repositório. No contexto do historiador, ao realizar uma pesquisa sobre determinado assunto, é de suma importância obter informações de diversas fontes para que esse profissional possa chegar a uma conclusão mais precisa, que represente os fatos ocorridos. O cruzamento de dados de diversas fontes é um processo de anotação, normalmente manual, custoso e muito suscetível à falhas, como erros nas anotações.

Perante o problema descrito, foram extraídos de um conjunto de documentos, as entidades nomeadas apresentadas na tabela 1 e suas respectivas relações. A importância desse processo se dá ao fato de que uma busca tem como objetivo encontrar alguma entidade ou sua característica. Os dados extraídos são representados através de um grafo que, posteriormente, é armazenado no Neo4j, um banco de dados específico para esse tipo de estrutura [Neo4j 2016]. Esse banco de dados disponibiliza uma série de funcionalidades que são exploradas pelo buscador para a realização de consultas.

Uma vez realizada a construção e o armazenamento do grafo, o sistema é capaz de realizar buscas considerando a forma na qual as entidades estão relacionadas. Para facilitar essas consultas, foi proposta uma nova linguagem que possui a seguinte sintaxe:

Entidade:“Especificação”

onde a entidade é a classe de entidades que se deseja buscar e a especificação é o atributo que caracteriza essa entidade, separados por dois pontos. Eventualmente, um dos elementos, entidade ou especificação, pode ser omitido. Para representar relações em entidades é utilizado o “- -”. Ele é utilizado em perguntas como: “Em quais datas Joaquim esteve relacionado a quais eventos?”. Essa pergunta, traduzida para a linguagem proposta seria:

Data - - Pessoa: “Joaquim” - - Evento

Espera-se que ao realizar uma busca, respeitando a sintaxe apresentada, o sistema apresente não só o documento que contém a informação, mas também em que parte do documento aquela informação se encontra. Será apresentado, portanto, uma lista de documentos onde ele pode ter acesso direto ao trecho onde se encontra a informação pesquisada; uma lista de entidades com algumas estatísticas; e por fim, será apresentado um grafo relacionando as entidades, provendo assim, uma interpretação visual dos dados.

5. Arquitetura do Sistema

O sistema proposto permite tanto buscas por palavras-chave quanto buscas semânticas. A figura 3 apresenta a arquitetura do sistema proposto. Inicialmente, dado um conjunto de documentos, é realizado um pré-processamento para que apenas dados textuais sejam extraídos e, posteriormente, esses dados são armazenados em um repositório. Para buscas por palavras-chave, é realizada apenas uma indexação tradicional. Já para as buscas semânticas, os documentos devem passar por um processo de extração de entidades nomeadas e suas relações, que então são usadas para a construção de um grafo. Em um banco de dados orientado a grafos, é armazenado o grafo construído na etapa anterior. Esse grafo também é utilizado para a realização da indexação semântica, uma vez que com ele sabe-se como as entidades se relacionam. Ao final das etapas informadas, os dois tipos de busca – por palavras-chave e semântica – são disponibilizadas no servidor, podendo assim atender as requisições vindas dos usuários.

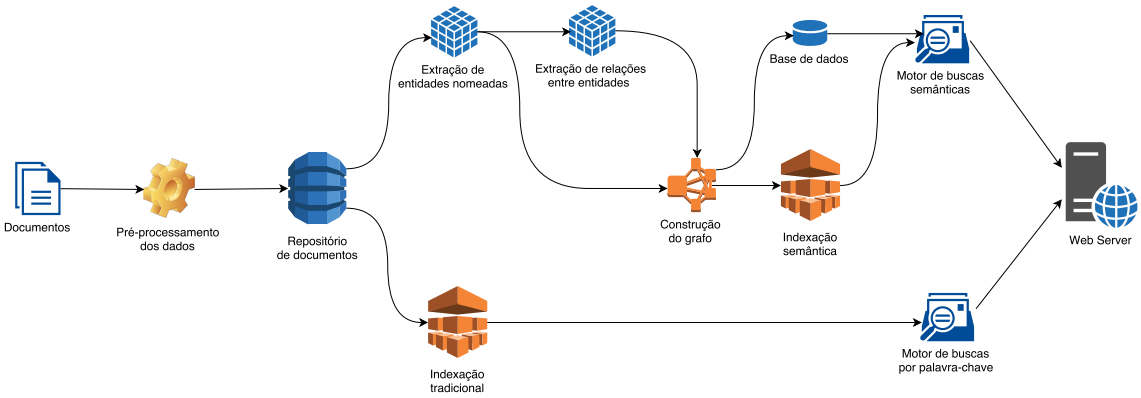


Figura 3. Arquitetura do buscador proposto.

6. Resultados

Os dados aqui utilizados são um conjunto de artigos relacionados à participação do exército brasileiro na Segunda Guerra Mundial. A principal fonte desses artigos é a revista Gazeta do Povo. Posteriormente, o sistema será alimentado diretamente pelo repositório do CEDIM.

Em consultas complexas, onde buscadores por palavras-chave apresentam dificuldades, o buscador semântico apresenta resultados de qualidade. Um exemplo desse cenário é quando o usuário busca obter a resposta da seguinte pergunta: "Quais pessoas compunham os grupos que participaram da guerra e quais são esses grupos?". Como os termos pessoas e grupos tratam-se de entidades nomeadas, um buscador comum não conseguiria encontrar exatamente o que o usuário deseja, uma vez que esses termos generalizam uma variedade de possibilidades. Já no buscador semântico proposto, esse problema seria facilmente resolvido com a seguinte consulta:

Pessoa - - Grupo - - Evento: "guerra"

Nessa consulta busca-se pessoas ligadas a grupos, que por sua vez está relacionado a um evento nomeado como "guerra". O resultado dessa busca é apresentado na forma de uma lista de documentos (Figura 4), lista de entidades (Figura 5) e um grafo (Figura 6).

Figura 4. Lista de documentos obtidos com a consulta.

Simple Results | Data Base Results | Graph Results

Entity	Slice	Citations	Relations	Document
Pessoa	Max Wolff	1	2	A-cobra-realmente-fumou-2-3.txt
Pessoa	Emílio Rodrigues Franklin	1	2	A-cobra-realmente-fumou-2-3.txt
Pessoa	Ari Schnaibel	1	3	A-cobra-realmente-fumou-2-3.txt
Grupo	nazista	2	3	A-cobra-realmente-fumou-2-3.txt
Pessoa	Reynaldo Pontarolli	1	1	A-cobra-realmente-fumou-2-3.txt
Grupo	pracinhas	9	13	A-cobra-realmente-fumou-2-3.txt
Evento	2.ª Guerra	3	5	A-cobra-realmente-fumou-2-3.txt
Pessoa	Modesto	2	2	A-cobra-realmente-fumou-2-3.txt
Pessoa	Nery	3	3	A-cobra-realmente-fumou-2-3.txt
Evento	2.ª Guerra Mundial	3	2	A-cobra-realmente-fumou-2-3.txt

Figura 5. Lista de entidades obtidas com a consulta.

Figura 6. Grafo obtido com a consulta.

7. Conclusões

Neste trabalho analisamos a necessidade que os historiadores possuem de um motor de buscas capaz de apresentar resultados de qualidade, mesmo para consultas complexas. Estudamos também as dificuldades enfrentadas por eles em cruzar dados de diversas fontes, assim como garantir uma margem aceitável de confiança sobre determinada informação.

Em parceria com o CEDIM, que apresenta um cenário como descrito, propomos um buscador semântico para o domínio da história. Sua principal característica perante aos buscadores convencionais caracteriza-se pela capacidade de análise de contexto da busca, trazendo assim, resultados mais precisos ao pesquisador. Outro fator relevante é que o processo de cruzamento de dados é realizado automaticamente, uma vez que o contexto é o principal critério de busca. Com o intuito de facilitar a interpretação e enriquecer os resultados, esses são apresentados de três formas diferentes, onde uma delas é a visualização gráfica, que fornece ao pesquisador a possibilidade de visualizar a maneira que as informações buscadas estão interligadas nos diversos documentos do repositório.

Além das funcionalidades apresentadas, buscaremos explorar outras linhas de pesquisa como: a recuperação de informação por conteúdo em mídias diversas, como áudio; a facilitação das buscas aproximando a sintaxe à linguagem natural; e a apresentação dos resultados por ordem de relevância utilizando técnicas de ranqueamento. Tais funcionalidades darão aos resultados ainda mais qualidade, além de abranger uma gama maior de dados, podendo assim, fornecer mais informações nas consultas.

Referências

CEDIM (2016). Centro de documentação e imagem. https://goo.gl/gfC3Xg, Agosto.

Cohen, S., Mamou, J., Kanza, Y., and Sagiv, Y. (2003). Xsearch: A semantic search engine for xml. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, pages 45–56. VLDB Endowment.

Guha, R. and McCool, R. (2003). Tap: A semantic web test-bed. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(1):81–87.

Guha, R., McCool, R., and Miller, E. (2003). Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, pages 700–709. ACM.

Haase, P., Broekstra, J., Ehrig, M., Menken, M., Mika, P., Olko, M., Plechawski, M., Pyszlak, P., Schnizler, B., Siebes, R., et al. (2004). Bibster–a semantics-based bibliographic peer-to-peer system. In *International Semantic Web Conference*, pages 122–136. Springer.

History-Lab (2016). Explore the archive. http://www.history-lab.org/overview, Agosto.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289.

Mäkelä, E. (2005). Survey of semantic search research. In *Proceedings of the seminar on knowledge management on the semantic web*. Department of Computer Science, University of Helsinki, Helsinki.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbís, J. M. (2013). Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Neo4j (2016). Neo4j: The world's leading graph database. https://neo4j.com/, Agosto.

Nobata, C., Sasaki, Y., Okazaki, N., Rupp, C., Tsujii, J., and Ananiadou, S. (2009). Semantic search on digital document repositories based on text mining results. In *International Conferences on Digital Libraries and the Semantic Web*, pages 34–48.

Ramachandran, A. and Sujatha, R. (2011). Semantic search engine: A survey. *International Journal of Computer Technology and Applications*, 2(6).

Renteria-Agualimpia, W., López-Pellicer, F. J., Muro-Medrano, P. R., Nogueras-Iso, J., and Zarazaga-Soria, F. J. (2010). Exploring the advances in semantic search engines. In *Distributed Computing and Artificial Intelligence*, pages 613–620. Springer.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

W3C (2014). Resource description framework (rdf). https://goo.gl/b3l3I9, Agosto.

Zhou, G. and Su, J. (2002). Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480. Association for Computational Linguistics.

Uma Aplicação Interligando Dados de GPS com Linked Geo Data

Gabriel de Sá Rodrigues[1], Gian Paixão[1], André Brito[1]

[1]Bacharelado em Ciência da Computação, Departamento de Ciência da Computação – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brazil

{gabrieldesar, gian.paixao, andre.brito}@ufrj.br

Abstract. *This article presents a short study about geolocated open data interlinked with Linked Geo Data database, specifically bus sensors datasets, using GeoSPARQL for spacial operations. From collection, through storage and data integration, the use benefits are analyzed and confronted with the challenges to disseminate this technology.*

Resumo. *Este artigo apresenta um breve estudo sobre interligação de dados abertos geolocalizados com a base do Linked Geo Data, especificamente dados de sensores de ônibus, utilizando GeoSPARQL para operações espaciais. Desde a coleta, passando pelo armazenamento e integração dos dados, são analisadas as vantagens da utilização dessas tecnologias, confrontando com os desafios para que seja mais difundida.*

1. Introdução

Após a ocorrência de um dos seus maiores desastres naturais de sua história em abril de 2010, em virtude das chuvas que resultaram 72 óbitos e milhares de desabrigados, foi criado no Rio de Janeiro uma estrutura para que a cidade respondesse com maior agilidade a situações de emergência, incluindo consequências de fenômenos naturais. A partir dessa iniciativa e de acordo com a Lei de Acesso à Informação[1], nasceu o Portal de Dados Abertos da Prefeitura do Rio de Janeiro (Data Rio)[2], que apresenta mais de 1.200 fontes de dados sobre, dentre outras categorias, saúde, educação, lazer e transporte.

Apesar dos esforços para tornar essas informações públicas, têm-se encontrado uma grande dificuldade para realizar a integração desses dados, tanto pela qualidade do conteúdo divulgado, quanto pela falta de um vocabulário comum entre as diversas entidades governamentais. Isso dificulta a utilização dessas bases para estudos e reduzem o potencial de sua utilização para encontrar soluções para os problemas da cidade.

Observando a importância de interligar dados, Tim Berners-Lee (2006) propôs um modelo de maturidade para a divulgação de dados abertos, começando desde a sua disponibilização através de um formato legível por máquina, e em um padrão aberto (como o CSV); passando pela disponibilização em um formato de *Linked Data*; para por fim interligá-lo a outro recurso existente.

Outras opiniões podem ser encontradas sobre importância desse tipo de interconexão de bases, com as de Lishan Zhang (2013), que afirma que dados interligados

[1] Decreto 7.724/2012
[2] http://data.rio/, acessado em outubro de 2016

auxiliam as análises através da expansão do banco de dados utilizando-se dados relacionados de outras fontes. Isso permite encontrar informações mais úteis e completas durante a realização de análises e estudos.

Através de uma análise anterior, pode-se observar que a maioria dos datasets divulgados pelo portal Data Rio estão em um padrão aberto, porém, não são divulgados num formato de *Linked Data*. Tendo isso em vista, nosso objetivo é, através de uma aplicação prática, verificar o processo de triplificação de uma base aberta do portal Data Rio e interligá-lo com uma base da *Web de Dados*, verificando as vantagens, possibilidades e desafios do processo.

2. Trabalhos Relacionados

A publicação do portal de dados abertos e disponibilização do *dataset* do setor de transportes incentivou diversos estudos na área. Matheus & Ribeiro (2014) analisam as causas e os desafios de implantação do portal de dados abertos. Marujo (2015) utilizou métodos de análise estatística para analisar o desempenho operacional do serviço, buscando gargalos e pontos críticos da frota. Bessa (2016) apresenta uma ferramenta que auxilia os usuários a detectarem o comportamento de ônibus fora de sua rota, utilizando um algoritmo de redes neurais.

Esses estudos demonstram o esforço que tem sido aplicado no tratamento e extração de conhecimento dessa base. Nosso trabalho contribuirá com a demonstração do cenário atual de implantação e utilização de *Linked Data* em dados abertos governamentais, mostrando as vantagens de ampliar o conhecimento através de interligações com outras bases públicas.

3. Fontes de Dados

O Data Rio disponibiliza uma API para consulta dos dados de GPS dos ônibus referentes ao último instante disponível em sua base, não fornecendo informações relacionadas a períodos anteriores. Dessa forma é preciso utilizar uma ferramenta de coleta e armazenagem desses dados para se obter uma quantidade de informação história suficiente que possibilite analisar o comportamento dos ônibus ao longo de um período de tempo determinado. Esses dados, fornecidos no formato JSON[3], possuem o identificador do ônibus, linha que opera, latitude, longitude, velocidade instantânea, além da data e hora da coleta.

O Linked Geo Data[4] é uma base de conhecimento em RDF[5] derivada do OpenStreetMaps[6], segundo os princípios de dados interligados. Essa base possui em seu acervo dados sobre lugares, pontos de interesse, rotas de transporte, dentre outras informações geolocalizadas. Com o foco de adicionar a dimensão espacial na *Web de Dados*, suas entidades possuem uma geometria e integração com a linguagem de consulta GeoSPARQL, facilitando operações espaciais. Em virtude disso, escolheu-se utilizar o

[3] JSON é acrônimo em inglês para JavaScript Object Notation, um formato de arquivo utilizado para a distribuição de dados.
[4] http://linkedgeodata.org/About, acessado em outubro de 2016
[5] RDF é o acrônimo em inglês para Resource Description Framework, uma linguagem utilizada para representar informação na internet.
[6] OpenStreetMap é um projeto colaborativo, no formato wiki, que permite o mapeamento de cidades ao redor do mundo.

Linked Geo Data para esse estudo, uma vez que há a possibilidade de integração com os dados de sensores dos ônibus.

4. Projeto de Dados Interligados

Durante o projeto analisamos quatro abordagens de consultas entre dados de diferentes bases: acompanhamento de consultas; utilização de uma base central; construir uma cópia local; ou utilizar um sistema de consultas federadas. A primeira abordagem visa estruturar a consulta entre diferentes *endpoints*, de maneira que a saída de uma consulta fornece a entrada para a base seguinte. Dessa forma é possível implementar uma lógica que automatize o processo de consulta a diferentes *datasets* através de seus *endpoints*. A segunda abordagem sugere a utilização de um *endpoint* que já possua a cópia de diferentes *datasets* incluída nele, como é o caso do *endpoint* da Open Link Software[7]. A terceira abordagem é semelhante à segunda, exceto pelo fato de que a infraestrutura e definição dos datasets utilizados é própria. A última abordagem propõe realizar a consulta através de um mediador, que se torna responsável por distribuir as subconsultas para as fontes relevantes e integrar os resultados.

No nosso estudo utilizamos a terceira abordagem pela facilidade de carga e gerência dos *datasets* a serem utilizados. Além disso, não era preciso um esforço de desenvolvimento em uma lógica específica. Também encontramos dificuldades em relação a complexidade de utilização das consultas federadas no banco de triplas utilizado. Também foi preciso desconsiderar a segunda abordagem, visto que estaríamos integrando um *dataset* ainda não publicado em formato de triplas e, portanto, não disponível em uma base central.

5. Interligação de Dados

Foi necessário dividir o processo em três fases: coleta de dados; transformação e carga; e realização das consultas. Para a fase de coleta dos dados dos ônibus, utilizamos o site da equipe do Laboratório de Engenharia de Software do Instituto de Computação da UFF (SEL-UFF/RJ)[8], tendo seus insumos catalogados por dia. A utilização do portal do SEL-UFF foi de fundamental importância para a obtenção de dados históricos, uma vez que fora realizado por eles um processo de extração e armazenamento dos dados do GPS dos ônibus obtidos através do Data Rio. Esses insumos foram tratados e carregados em um banco de dados relacional. Em relação aos dados do Linked Geo Data, utilizamos a base disponível no site oficial do projeto[9].

Na fase de transformação e carga, utilizamos a ferramenta D2RQ para gerar o mapeamento do banco de dados relacional para o formato de triplas. Após a geração, definimos o namespace a ser utilizado na URI e geramos a extração RDF, serializado em Turle para ser carregado num banco de triplas. Utilizamos a ferramenta Virtuoso OpenSource 6.1 para armazenar as triplas a serem interligadas, através da funcionalidade Quad Store Upload, conforme figura 1.

[7] http://lod.openlinksw.com/sparql, acessado em outubro de 2016
[8] http://sel.ic.uff.br, acessado em abril de 2016
[9] http://downloads.linkedgeodata.org/releases/, acessado em outubro de 2016

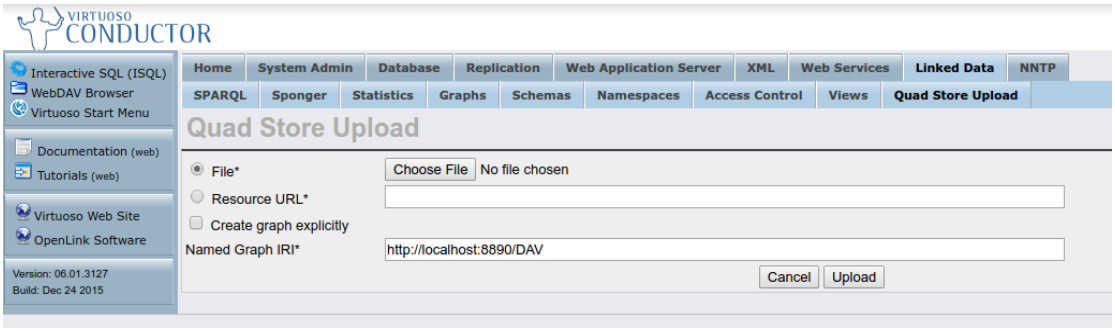


Figura 1. Tela de carregamento de bases RDF no banco Virtuoso Opensource 6.1

Na última fase foi preciso definir as consultas em SPARQL a serem utilizadas. Para aproveitar as características dos dados geolocalizados, utilizamos funções de GeoSPARQL de interseção, de forma a ser possível, através de um par latitude e longitude de um dado ônibus no tempo, identificar pontos de interesse e locais na sua proximidade. É possível passar como parâmetro o raio de interseção, em quilômetros.

Dessa forma, construímos duas possíveis consultas, uma retornando pontos de interesse ao redor de um dado ônibus e outra retornando à referência de lugar mais próxima.

Tabela 1. Consulta em SPARQL de busca por pontos de interesse na proximidade de 100 metros

```
Prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
Prefix ogc: <http://www.opengis.net/ont/geosparql#>
Prefix geom: <http://geovocab.org/geometry#>
Prefix lgdo: <http://linkedgeodata.org/ontology/>
Prefix datario: <http://datario/bus_rdf/>

SELECT ?linha ?amenity
WHERE {
    ?s a lgdo:Amenity ;
    rdfs:label ?amenity ;
    geom:geometry [
      ogc:asWKT ?g
    ] .

    ?busp a datario:BusPosition ;
    rdfs:label ?linha ;
    datario:longitude a long? ;
    datario:latitude a lat? ;

  Filter (
    bif:st_intersects (?g, bif:st_point (?long, ?lat), 0.1)
  ) .
}
```

Tabela 2. Consulta em SPARQL de busca por local atual através de ordenamento entre locais mais próximos

```
Prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
Prefix ogc: <http://www.opengis.net/ont/geosparql#>
Prefix geom: <http://geovocab.org/geometry#>
Prefix lgdo: <http://linkedgeodata.org/ontology/>
Prefix datario: <http://datario/bus_rdf/>

SELECT ?linha ?place
WHERE {
    ?s a lgdo:Place ;
    rdfs:label ?place ;
    geom:geometry [
      ogc:asWKT ?g
    ] .

    ?busp a datario:BusPosition ;
    rdfs:label ?linha ;
    datario:longitude a long? ;
    datario:latitude a lat? ;

  Filter (
    BIND (STRDT(CONCAT("POINT(",?long, " ", ?lat, ")"),ogc:WktLiteral) as ?p)
  ) .
}
ORDER BY ASC(ogc:distance(?g, ?p)) LIMIT 1
```

6. Conclusão

Essa aplicação evidenciou que, apesar de existir ferramental para coletar, armazenar e consultar os dados, foi necessário utilizar diversas ferramentas para conseguir alcançar o objetivo final. Além disso, tivemos o desafio de contornar limitações e problemas apresentados por essas ferramentas, visto não estar disponível outra opção que integrasse todo o processo de maneira mais simplificada.

A utilização de GeoSPARQL foi adequada para os *datasets* utilizados, representando um ganho pela possibilidade de utilização de operações espaciais, não sendo necessário referenciar objetos estaticamente pelo nome, mas sim através de sua posição geoespacial. É necessário um estudo mais aprofundado das funções oferecidas, aplicação em outros domínios que não o de transportes e integração com outros tipos de dados geolocalizados, de forma a ser possível avaliar a sua robustez e aplicação.

Projetar a arquitetura de dados interligados se mostrou uma parte importante do processo, visto que não só altera as consultas realizadas, mas impacta no desempenho e atualidade dos dados. A opção escolhida serviu pela abordagem exploratória do estudo, mas foram encontradas dificuldades na necessidade de coleta contínua e armazenagem em um servidor central. É importante avaliar a opção de consultas federadas, no âmbito de aplicação e de desempenho, visto sua natureza descentralizada que se adequa mais à proposta de *Linked Data*.

Essas dificuldades se aplicam principalmente à esfera governamental, provavelmente em virtude da dependência de um ferramental e técnicas já ultrapassadas no mercado para a divulgação dos dados. Essas ferramentas, por mais consolidadas que estejam, não atendem ao nível de maturidade que se espera para a interligação dos dados, tornando-os difíceis de se manipular. Espera-se que, de maneira gradual, as diversas

autarquias governamentais comecem a evoluir seus ambientes de publicação e exploração de dados, passando a utilizar maior padronização na divulgação dos dados, e aumentando a qualidade na descrição e apresentação dessa informação para a população.

Esperamos que futuramente possamos divulgar mais amplamente os resultados obtidos através dessas análises, bem como verificar as implicações de utilizar um maior número de fontes de dados integradas para gerar informações úteis aos cidadãos.

Agradecimentos

Agradecemos à professora Maria Luiza Campos pela orientação durante o desenvolvimento deste artigo e revisão. Também agradecemos à equipe do Laboratório de Engenharia de Software do Instituto de Computação da UFF (SEL-UFF/RJ) por gentilmente ter disponibilizado suas bases para o público em geral.

Referências

BERNERS-LEE, T. **Linked Data**. 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 14 out. 2016.

BESSA, A. et al. **RioBusData: Outlier Detection in Bus Routes of Rio de Janeiro**. 2007.

HART, G.; DOLBEAR, C. **Linked Data: A Geographic Perspective**. [s.l.] CRC Press, 2013.

HARTIG, O. **Querying Linked Data with SPARQL**. 2009. Disponível em: <http://pt.slideshare.net/olafhartig/querying-linked-data-with-sparql>. Acesso em: 14 out. 2016.

KOLAS, D.; PERRY, M.; HERRING, J. **Getting Started with GeoSPARQL**. 2013. Disponível em: <http://www.ssec.wisc.edu/meetings/geosp_sem/presentations/GeoSPARQL_Getting_Started - KolasWorkshop Version.pdf>. Acesso em: 14 out. 2016.

MARUJO, L. G. et al. **Um método para avaliação do desempenho de ônibus baseado em dados de GPS.** XXIX Congresso Nacional de Pesquisa em Transportes da ANPET, p. 1194–1205, 2015.

MATHEUS, R.; RIBEIRO, M. M. **Case Study Open Government Data in Rio de Janeiro City**. p. 1–50, 2014.

OPEN LINK SOFTWARE. **Geometric Objects**. 2014. Disponível em: <http://docs.openlinksw.com/virtuoso/sqlrefgeospatialgo/>. Acesso em: 14 out. 2016.

ZHANG, L. **How structured data (Linked Data) help in Big Data Analysis---Expand Patent Data with Linked Data Cloud**. 2013.

Utilização de Sistema Especialista para Diagnósticos de Doenças Transmitidas pelo Aedes Aegypti

Vitor de L. O. Fonseca[1], Luiz H. S. Volpasso[2], Gizelle K. Vianna[3]

[1,2,3]Instituto de Ciências Exatas – Departamento de Matemática – Universidade Federal Rural do Rio de Janeiro (UFRRJ)

BR 465 – Km 7 – CEP 23890-000 – Seropédica – RJ – Brasil

{vitorlofonseca, kupac}@ufrrj.br[13], luizvolpasso@pet-si.ufrrj.br[2]

Abstract. *In the last years, the african Aedes Aegypti mosquito has made many victims in Brazil and in countries around the Equator, serving as a vector for three known viruses: Dengue, Chikungunya and Zika. Moreover, the corrected diagnose of the viruses is an issue, as the symptoms can be mixed. This work presents an alternative way to diagnose patients that host one of the three viruses, by using an expert system that analyzes the combined symptoms and displays a probabilistic diagnosis.*

Resumo. *Nos últimos anos o mosquito Aedes Aegypti, de origem africana, tem feito muitas vítimas no Brasil e em países próximos a linha do Equador, servindo de vetor para três vírus conhecidos: Dengue, Zika e Chikungunya. Além da existência do Aedes, existe um grande problema relacionado a esses três vírus, que é o diagnóstico diferencial correto entre as três doenças. Esse trabalho apresenta uma forma alternativa de diagnosticar pacientes, através da análise dos sintomas apresentados por um sistema especialista que emite um diagnóstico probabilístico da doença.*

1. Introdução

Em 2016, foram registrados 1.426.005 casos prováveis de dengue no país até a Semana Epidemiológica 32 (3/1/2016 a 13/08/2016). No mesmo período do ano de 2015, foram estimados 1.479.950 casos (Figura 1). Nessas 32 primeiras semanas do ano de 2016, houveram 509 mortes confirmadas pelo Ministério da Saúde. No mesmo período, estima-se que 216.102 pessoas foram infectadas pelo vírus Chikungunya, e 196.976 pelo vírus Zika [Ministério da Saúde 2016]. Diversas medidas para a erradicação dos vírus citados anteriormente já foram tomadas por instituições governamentais, como distribuição de insumos estratégicos, como inseticidas e kits para diagnóstico aos estados e municípios, elaboração e disponibilização de cursos virtuais relacionados ao assunto [Ministério da Saúde, 2016], passagens frequentes de carros fumacê para o lançamento de inseticida, campanhas de prevenção e combate na internet, entre outras.

Porém um dos problemas que mais tem afetado os órgãos de saúde nos últimos 2 anos é identificar qual dos três vírus um paciente está hospedando. Isso acontece porque os sintomas dos três são quase idênticos (Tabela 1). Em decorrência disto, há a necessidade de que o médico utilize uma ponderação adequada para cada tipo de sintoma que o paciente possuir, visto que, ajudará na diferenciação de qual doença o paciente possui.

Além disso, tem-se dificuldade na compreensão dos médicos nas descrições dos sintomas que o paciente tem, por não serem precisas ou completas, e também por perguntas mal formuladas realizadas pelos médicos. Estes fatores tornam a diagnóstico mais complexo.

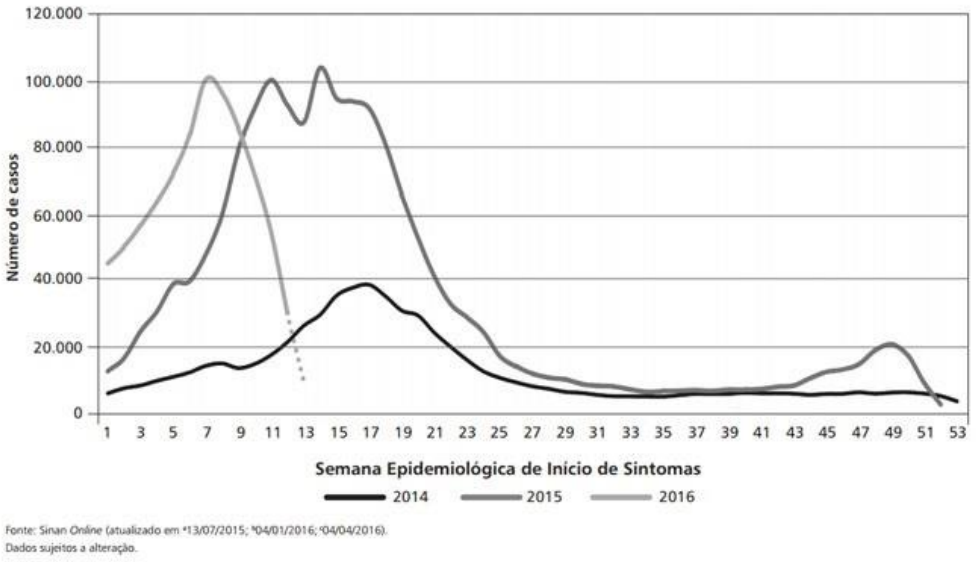


Figura 1. Casos prováveis de Dengue por semana no ano de 2016, 2015 e 2014 (Secretaria de Vigilância em Saúde - Ministério da Saúde, 2016).

Tabela 1. Sintomas e seus respectivos vírus.

Sinais/Sintomas	Dengue	Zika	Chikungunya
Febre (duração)	Acima de 38 ºC (4 a 7 dias)	Sem febre ou subfebril 38 ºC (1-2 dias subfebril)	Febre alta > 38 ºC (2-3 dias)
Manchas na pele (frequência)	A partir do 4º dia (30-50% dos casos)	Surge no 1º ou 2º dia (90-100% dos casos)	Surge 2-5 dias (50% dos casos)
Dor nos músculos (frequência)	+++/+++	++/+++	+/+++
Dor na articulação (frequência)	+/+++	++/+++	+++/+++
Intensidade da dor articular	Leve	Leve/Moderada	Moderada/Intensa
Edema na Articulação	Raro	Frequente e leve intensidade	Frequente e de moderada a intensa
Conjuntivite	Raro	50-90% dos casos	30%
Dor de cabeça (frequência e intensidade)	+++	++	++
Coceira	Leve	Moderada/Intensa	Leve
Alteração no sistema nervoso	Raro	Mais frequente que Dengue e Chikungunya	Raro (predominante em Neonatos)

O objetivo desse trabalho é auxiliar no diagnóstico dos pacientes, utilizando as frequências e intensidades dos sintomas dos hospedeiros como entradas de um sistema especialista, construído através do Expert Sinta [SINTA 2016], que utiliza um modelo de representação do conhecimento baseado em regras de produção e probabilidades. Deve-

se salientar que o sistema apenas auxilia na elaboração do diagnóstico, não pretendendo ser categórico em um diagnóstico final.

2. Processo de Diagnóstico através de um Sistema Especialista

Um sistema especialista consiste em um sistema lógico composto por um conjunto de regras de produção. As regras fornecem uma forma de representação do conhecimento bastante próxima da forma humana de expressá-lo, e podem ser estruturadas como sentenças booleanas ou binárias do tipo SE...ENTÃO...SENÃO, armazenadas em uma base de conhecimento [LOPES, 2005]. Formando um conjunto de regras e fatos, pode-se extrair um determinado estado conclusivo (objetivo) que, em aplicações médicas, seria um diagnóstico.

Abaixo, mostraremos o conjunto de regras usadas para representar os fatos conhecidos sobre o diagnóstico das doenças em questão e a forma usada para ponderar os fatos em busca de um objetivo, com suas respectivas conclusões.

2.1. Obtenção de conclusões com operadores lógicos e fatos

Nesse trabalho, utilizamos abordagem orientada a objetivos, a qual aciona as regras que contém os objetivos atuais, na ordem em que se apresentam. A cada regra acionada, caso a premissa não possa ser avaliada, novos objetivos são definidos e o processo de avaliação das regras é reiniciado.

Para que o processo de obtenção de objetivos se assemelhe mais às decisões de especialistas humanos, sistemas especialistas utilizam o artificio da probabilidade. Com a utilização desse recurso, regras, premissas e conclusões são ponderados por fatores de confiança, que consistem de valores percentuais que representam a confiabilidade de cada um desses elementos. Fatores de confiança devem corresponder a um número entre 0% e 100%, proporcionalmente à expectativa de que aquele elemento seja verdadeiro. Define-se também um limiar, onde se o fator de confiança de determinada afirmativa for menor que esse limiar, tal afirmativa é considerada falsa, do contrário verdadeira.

Neste trabalho, foram usados três formatos de regra de produção. O primeiro formato é a implicação básica na forma SE A ENTÃO B onde, se a condição A for verdadeira, a conclusão B também o será e falsa, caso contrário. No segundo formato aparecem as implicações com conjunções, do tipo SE A e B ENTÃO C. Nesse tipo de regra, apenas quando A e B forem condições verdadeiras simultaneamente, C será uma conclusão verdadeira. Se A ou B forem falsas, C será falsa. No terceiro formato, temos as implicações disjuntivas SE A ou B ENTÃO C, onde C só será falso se A e B forem falsos ao mesmo tempo. O primeiro passo de representação do conhecimento é identificar o formato da implicação e, a partir daí, temos três formas para calcular o fator de confiança do objetivo da mesma, de acordo com o formato de regra sendo usado, a saber:

1) Para regras do tipo SE A então B, a forma de calcular o fator de confiança de um objetivo (B) consiste em multiplicar o fator de confiança da premissa (A) pelo fator de confiança da regra.
2) Para regras do tipo SE A e B então C, multiplica-se o fator de confiança das premissas (A) e (B), pelo fator de confiança da regra.

3) Para regras do tipo SE A ou B então C, devemos subtrair o produto dos fatores de confiança das premissas (A) e (B), da soma desses mesmos fatores e multiplicar o resultado pelo fator de confiança da regra.

2.2. Listagem e descrição das regras

Ao todo, foram listadas e utilizadas 34 regras com bases nos sintomas relacionados a cada uma das doenças (Tabela 2).

Tabela 2. Sintomas Principais e Regras

REGRA 1	SE temperatura corporal = acima de 38 ºC E duração febre = 2 a 3 dias	ENTÃO chikungunya
REGRA 2	SE temperatura corporal = acima de 38 ºC E duração febre = 4 a 7 dias	ENTÃO dengue
REGRA 3	SE temperatura corporal = abaixo de 38 ºC	ENTÃO zika
REGRA 4	SE dia de aparecimento de manchas na pele = 1º ou 2º	ENTÃO zika
REGRA 5	SE dia de aparecimento de manchas na pele = 2º ao 5º	ENTÃO chikungunya
REGRA 6	SE dia de aparecimento de manchas na pele = a partir do 4º	ENTÃO dengue
REGRA 7	SE dia de aparecimento de manchas na pele = não tem manchas	ENTÃO dengue
REGRA 8	SE dor nos músculos = com dor E intensidade = intensa	ENTÃO dengue
REGRA 9	SE dor nos músculos = com dor E intensidade = moderada	ENTÃO zika
REGRA 10	SE dor nos músculos = com dor E intensidade = leve	ENTÃO chikungunya
REGRA 11	SE dor nos músculos = sem dor	ENTÃO não tem doença
REGRA 12	SE dor nas articulações = com dor E intensidade = intensa	ENTÃO chikungunya
REGRA 13	SE dor nas articulações = sem dor	ENTÃO não tem doença
REGRA 14	SE dor nas articulações = com dor E intensidade = leve	ENTÃO dengue
REGRA 15	SE dor nas articulações = com dor E intensidade = moderada	ENTÃO zika
REGRA 16	SE edema nas articulações = não tenho	ENTÃO dengue
REGRA 17	SE edema nas articulações = leve intensidade	ENTÃO zika
REGRA 18	SE edema nas articulações = moderada ou intensa	ENTÃO chikungunya
REGRA 19	SE conjuntivite = tenho	ENTÃO zika e chikungunya
REGRA 20	SE conjuntivite = não tenho E intensidade nas dores articulares = intensa	ENTÃO chikungunya
REGRA 21	SE conjuntivite = não tenho E intensidade nas dores articulares = leve	ENTÃO dengue
REGRA 22	SE dor de cabeça = sem dor	ENTÃO não tem doença
REGRA 23	SE dor de cabeça = com dor E intensidade = leve	ENTÃO chikungunya e zika
REGRA 24	SE dor de cabeça = com dor E intensidade = moderada E intensidade nas dores articulares = intensa	ENTÃO chikungunya
REGRA 25	SE dor de cabeça = com dor E intensidade = intensa	ENTÃO dengue
REGRA 26	SE dor de cabeça = com dor E intensidade = moderada E intensidade nas dores articulares = moderada	ENTÃO zika
REGRA 27	SE coceira = não tenho	ENTÃO não tem doença

REGRA 28	SE coceira = tenho E intensidade coceira = leve E intensidade das dores musculares = intensa	ENTÃO dengue
REGRA 29	SE coceira = tenho E intensidade coceira = leve E intensidade das dores musculares = leve	ENTÃO chikungunya
REGRA 30	SE coceira = tenho E intensidade coceira = moderada	ENTÃO zika
REGRA 31	SE coceira = tenho E intensidade coceira = intensa	ENTÃO zika
REGRA 32	SE alterações no sistema nervoso = sim	ENTÃO zika
REGRA 33	SE alterações no sistema nervoso = não E intensidade nas dores articulares = leve	ENTÃO dengue
REGRA 34	SE alterações no sistema nervoso = não E intensidade nas dores articulares = intensa	ENTÃO chikungunya

Após a formulação das regras com suas respectivas condições, basta inicializar a aplicação que o programa realizará o questionário e, com base nas respostas dados pelo usuário, encontrará a probabilidade de classificação em cada uma das doenças (objetivos) que o usuário possa ter.

O programa não necessariamente acionará todas as regras ou passará por todas elas pois, conforme ele vai sendo alimentado de informações, passa a ter uma inferência individual com base nos questionários respondidos por cada paciente e pelas decisões tomadas durante o percurso realizado na árvore de pesquisa. Além disso, o programa pode realizar novas buscas sempre que um novo objetivo for definido, quando uma regra acionada não puder ser respondida sem que novos fatos sejam concluídos.

Ao final do questionário é informado ao usuário a probabilidade de diagnóstico para cada uma das três doenças avaliadas, além de disponibilizar o histórico do processamento de todas as regras avaliadas para aquele caso individual (Figura 2). Lembrando que existe um limiar de corte que pode desprezar regras e conclusões, nem sempre todas as doenças terão um valor significativo de probabilidade para que possam ser consideradas como possíveis de existir em um determinado quadro.

3. Conclusões

Os Sistemas Especialistas se mostram capazes de atuar em diferentes áreas de aplicação, com bastante flexibilidade quanto ao número e formato das regras necessárias para modelar um ramo do conhecimento. Para isso, porém, o sucesso depende que o responsável pela tradução do conhecimento para o formato de conjunto de regras lógica tenha uma boa capacidade de organização estruturada das informações fornecidas por especialistas das áreas e das conexões entre as partes desse conhecimento.

No estudo de caso aqui apresentado, abordamos a adequação do paradigma ao diagnóstico das doenças relacionadas ao vetor Aedes que são, em nível de sintomas, praticamente idênticas. Utilizando o sistema da forma correta, com um especialista que tenha habilidade de agregar conhecimento extra quando necessário, haverá grande probabilidade do sistema atender os requisitos para o qual foi construído, podendo auxiliar setores de triagem em postos de emergência, por exemplo.

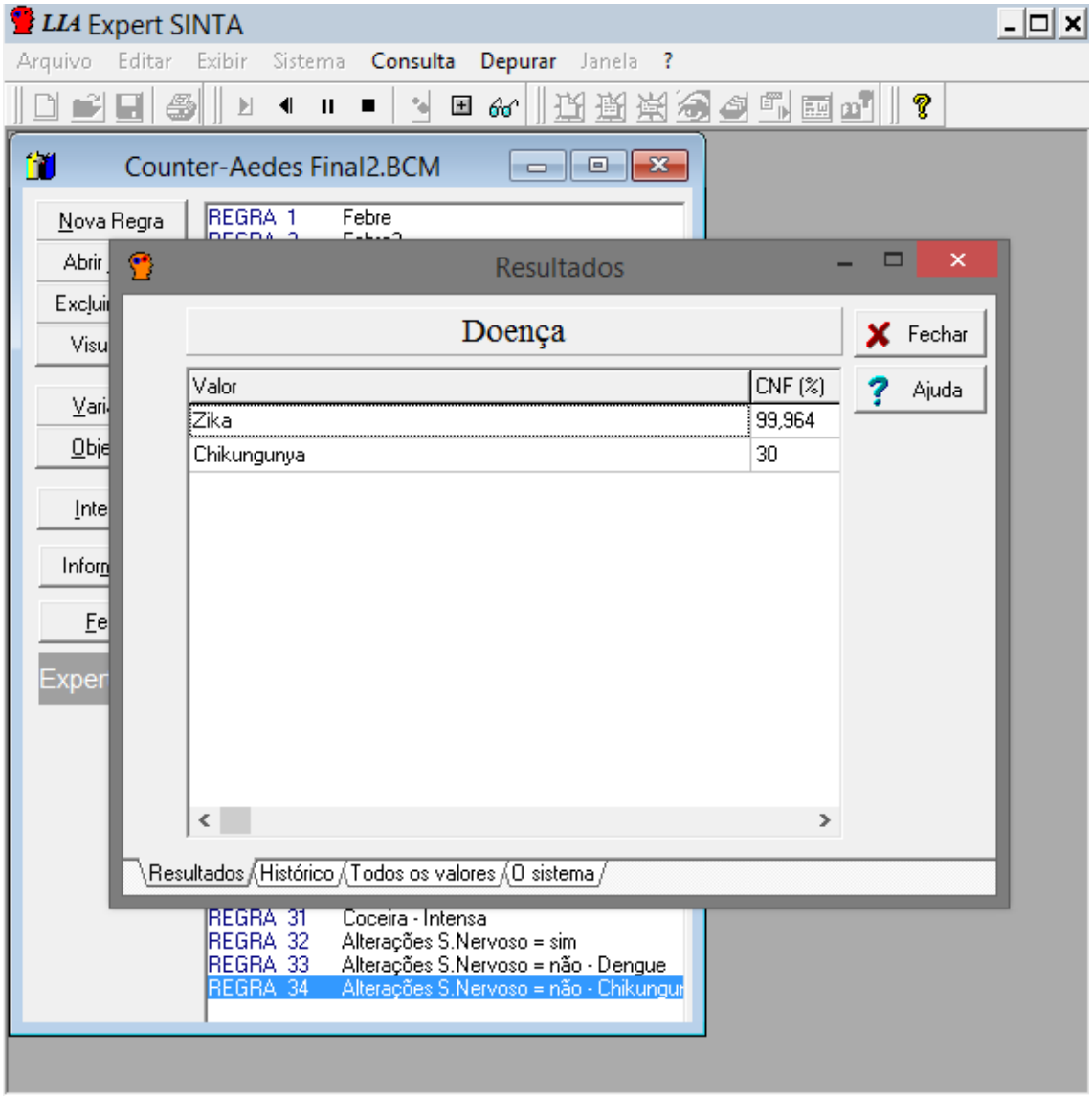


Figura 2. Resultado final após questionário

4. Referências

MS, (2016), "Boletim Epidemológico, Volume 47, Número 8", Secretaria de Vigilância em Saúde - Ministério da Saúde, página: http://www.combateaedes.saude.gov.br/images/pdf/2016-006-Dengue-SE5.pdf

LOPES, M.H.B. de Moraes, HIGA, R. (2005) "Desenvolvimento de um sistema especialista para identificação de diagnósticos de enfermagem relacionados com a eliminação urinária"

SINTA (2016), Projeto Expert Sinta, Laboratório de Pesquisa em Ciência da Computação, Departamento de Computação, UFC, página: http://www.lia.ufc.br/~bezerra/exsinta/ (Consultada em maio de 2016)

Uma abordagem algorítmica para auxiliar precocemente ao diagnóstico de jovens em risco de TDAH

Yara de Lima Araújo[1], José Raimundo Macário Costa[2], Sérgio Manuel Serra da Cruz[3]

[1,2,3] Instituto de Ciências Exatas - Departamento de Matemática– Universidade Federal Rural do Rio de Janeiro (UFRRJ)

BR-465, Km 7 - CEP 23.897-000 - Seropédica- RJ – Brasil

{yara,serra}@pet-si.ufrrj.br[1 3], mac_costa@yahoo.com[2]

Abstract. *Given the concern about academic, social and psychologic consequences in those diagnosed with ADHD, there is an opportunity to develop a tool for helping early diagnosis of ADHD. For this reason, computational techniques such as Neural Networks and Clustering algorithms are seen as useful alternatives by researchers. This paper presents results from tests with these algorithms. Within the study, it was possible to identify one possible candidate at risk of ADHD.*

Resumo. *Diante da preocupação com as consequências acadêmicas, sociais e psicológicas dos portadores do Transtorno de Deficit de Atenção e Hiperatividade, percebe-se que há oportunidades de desenvolvimento de ferramenta computacional capaz de auxiliar no diagnóstico precoce do TDAH. Para tal, técnicas como Redes Neurais e algoritmos de clusterização são vistas como boas alternativas pela comunidade de pesquisa. Este trabalho apresenta resultados da adoção destes algoritmos, e foi possível identificar um candidato em risco de TDAH.*

1. Introdução

Desde o estabelecimento das políticas de inclusão de portadores de necessidades especiais nas escolas regulares através da LDB 9394/96, artigo 58 (Brasil, 1996), profissionais da educação buscam orientações e informações acerca de como lidar com determinados transtornos de aprendizagem e ao mesmo tempo garantir o direito ao aprendizado a todos os alunos. Além disso, devido a grande ocorrência desses transtornos entre crianças e jovens, um grande número de pesquisadores de áreas interdisciplinares vem investigando o problema, buscando desenvolver aplicações computacionais para detectar precocemente os indivíduos com dificuldades de aprendizagem.

De acordo com o Manual Diagnóstico e Estatístico da Associação Americana de Psiquiatria (DSM-IV, 2014), um transtorno específico de aprendizagem, como o nome implica, é diagnosticado diante de déficits específicos na capacidade individual para perceber ou processar informações com eficiência e precisão. Dentre esses transtornos de aprendizagem podemos destacar o Transtorno de Deficit de Atenção e Hiperatividade (TDAH) manifesto inicialmente durante os primeiros anos de escolaridade formal; caracterizando-se por dificuldades persistentes nas habilidades básicas de leitura, escrita e/ou matemática (SAMPAIO e FREITAS, 2014, p. 131).

Essas características afetam o desenvolvimento acadêmico, os relacionamentos familiares e sociais e a vida laboral. O TDAH se manifesta na primeira infância e se não for corretamente diagnosticado e tratado precocemente pode apresentar sintomas na vida adulta, interferindo na vida acadêmica, profissional, afetiva e social. Estima-se que 70% das pessoas que tiveram TDAH diagnosticado na infância mantêm o transtorno na vida adulta (BASTOS et al. 2012).

Neste trabalho, investigamos a adoção de técnicas baseadas na utilização de algoritmos de clusterização que permitam a detecção precoce e automatizada de possíveis candidatos em risco de TDAH. Nossa pesquisa se justifica em estudos recentes que apontam de 3 a 5% de crianças atingidas por este transtorno (APA, 2002; ROHDE & KETZER, 1997). Além disso, ainda há desconhecimento sobre o TDAH mesmo por parte de pais e professores, o que pode resultar na identificação tardia do transtorno.

O objetivo desse trabalho é avaliar a eficácia de algoritmos de clusterização *K-means* e uma rede neural não supervisionada para a detecção de possíveis candidatos em risco de TDAH. Nesse trabalho foram extraídos dados de entrevistas realizadas com jovens brasileiros e espanhóis da faixa etária entre 8 a 18 anos através do questionário MTA-SNAP-IV.

Esse trabalho está organizado da seguinte forma: a seção 2 apresenta o referencial teórico e trabalhos relacionados. A seção 3 descreve os materiais e métodos utilizados e a amostra de dados coletados. Os resultados dos testes são apresentados na seção 4, seguida da discussão na seção 5. Finalmente, a seção 6 apresenta as considerações finais e possíveis trabalhos futuros.

2. Trabalhos Relacionados

Atualmente, existem métodos acessíveis para identificação do TDAH para pais e professores ou ao próprio portador do transtorno. Dentre eles, estão os questionários de escala que podem ser respondidos por pais, cuidadores e/ou professores. Destacamos o *ADHD Rating Scale*, o questionário de Conners e o SNAP-III e IV (MENDONÇA DE ANDRADE et al, 2011). Ressaltamos que os questionários fazem parte da primeira etapa do processo de diagnóstico do TDAH e que deve ser seguida de análise clínica por profissional de saúde especializado, ou seja, o caráter do questionário é de indicação e não de diagnóstico.

O TDAH, diferentemente do transtorno da dislexia, ainda carece de ferramentas computacionais voltadas para investigação precoce automatizada (COSTA *et al*, 2014). Nas investigações sobre dislexia, os autores utilizaram redes neurais para identificar possíveis portadores de dislexia. Nestes estudos, as redes neurais tiveram acertabilidade de aproximadamente 80% dos casos, o que foi considerado satisfatório.

Outro trabalho relacionado foi apresentado por (MIRANDA *et al*, 2011), onde os autores aplicaram o questionário MTA-SNAP-IV em um grupo de crianças com o objetivo de testar a eficácia do questionário. Foi possível indicar os sintomas presentes na amostra de crianças que poderiam levar ao diagnóstico de TDAH.

Tenev et al (2013) utilizaram técnicas computacionais não-supervisionada do tipo SVM para classificar sub-tipos de TDAH em adultos, através do eletroencefalograma em

diferentes situações. Os dados procedem de 67 pessoas previamente diagnosticadas com TDAH e outras 50 não-portadoras de qualquer transtorno neurológico. Os padrões foram definidos através da SVM por intermédio da lógica de Karnot produzindo a saída dos classificadores. Os resultados mostraram que mais de 80% das instâncias foram classificadas corretamente.

Neste sentido, nossa contribuição no presente trabalho é acelerar o processo de diagnóstico precoce de TDAH em crianças e jovens em idade escolar, fazendo uso de diferentes técnicas computacionais e algoritmos que se mostram interessantes métodos para classificação de dados.

3. Material e Métodos

O questionário utilizado nessa pesquisa foi o MTA-SNAP-IV, livremente disponibilizado na rede com tradução em Português (Mattos *et al.*, 2006). O questionário é composto por 26 perguntas das quais 18 relacionadas ao TDAH e 8 relacionadas ao Transtorno de Oposição. Das 18 perguntas, 9 são a respeito da hiperatividade e impulsividade, e as outras 9 sobre desatenção. Além destas, foram adicionadas 9 perguntas relacionadas ao comportamento de indivíduos de forma geral para fins de teste.

O questionário é aplicado a pais e professores, o que é importante, pois para diagnóstico de TDAH o indivíduo precisa ser analisado em relação a mais de um ambiente no qual está inserido. Além disso, como o desempenho acadêmico da criança portadora de TDAH é geralmente afetado, os professores tendem a perceber certos sintomas da criança no seu dia-a-dia.

3.1. Amostra de dados

O público selecionado foi composto por jovens da faixa etária entre 8 a 18 anos residentes na cidade do Rio de Janeiro, Brasil e em Salamanca, Espanha. Foram avaliados um total de 52 crianças/adolescentes. A tabela a seguir ilustra a distribuição das faixas etárias dos entrevistados.

Tabela 1. Idade da população alvo, adaptada de COSTA, 2011.

Faixa etária	Total por faixa etária	%
9-12	16	30
13-16	15	30
16 ou mais	21	40
Total	**52**	**100**

3.2. Tipo de pesquisa

Esta pesquisa tem caráter experimental qualitativo. De acordo com Wazlawick (2014, p. 23), este tipo de pesquisa implica ter uma ou mais variáveis experimentais que podem ser controladas pelo pesquisador, e uma ou mais variáveis observadas, cuja medição poderá levar à conclusão de que existe algum tipo de dependência com a variável experimental. Ou seja, dados provenientes dos entrevistados foram manipulados e analisados por ambos o algoritmo de clusterização *K-means* e a rede neural não supervisionada do tipo SOM, e a partir deste processo, é possível extrair informações como a influência de determinadas variáveis atuam no resultado obtido.

A forma de coleta de dados se deu através de questionário de escala, então foi necessario realizar um processo de quantificação para que os qualificadores tivessem compatibilidade com os requisitos do algoritmo. O questionário possui os qualificadores "*nem um pouco"*, "*pouco"*, "*bastante"*, "*demais*." Estes foram mapeados em números correspondentes (0,1,2,3) em um banco de dados para então servirem de input para o *k-means* e a rede neural em questão (SOM).

3.3. Algoritmo de clusterização *k-means* e rede neural não-supervisionada

O *k-means* é um algoritmo clássico relativamente simples, trabalha agrupando objetos em k números de grupos, com k >1. O número de grupos k que se deseja encontrar nos dados selecionados deve ser definido previamente pelo usuário (JAIN, MURTY e FLYNN, 1999; KANUNGO, et al, 2002). O algoritmo calcula os centros de um grupo (centróides). Neste trabalho, optamos por 3 clusters (k=3), pois foi o número mínimo necessário para assegurar a separação das variáveis em clusters distintos.

Nos algoritmos de aprendizado não-supervisionado não existe o papel do especialista treinando a rede. Neste tipo de rede, são disponíveis apenas padrões de entrada de treinamento de classificação desconhecida. O algoritmo avalia os conjuntos de dados apresentados, determina algumas propriedades dos conjuntos de dados e aprende estas propriedades na sua saída. O método de agrupamento de cada padrão segue um critério de similaridade e depende do algoritmo empregado, dos dados utilizados e da medida de similaridade adotada (CARVALHO, 2005).

Neste trabalho, optamos por utilizar algoritmo de rede neural não-supervisionada do tipo *SOM* (*Mapas Auto-Organizáveis*), também conhecidas como *Mapas de Kohonen*. As redes SOM utilizam o algoritmo "vencedor-leva-tudo", o que significa que apenas um neurônio com maior número de ativação irá fornecer a saída da rede, e somente seus pesos são alterados. Durante o chamado treinamento competitivo, o neurônio cujo peso for mais próximo do vetor de entrada será o vencedor, e esta proximidade é calculada como *distância euclidiana*. Para tal, a seguinte fórmula é utilizada:

$$\mathrm{di} = \sum_{j=1}^{n} (\mathrm{Wij} - \mathrm{Xj})^2$$

onde *n* é o número de neurônios de entrada (elementos vetor de entrada). A rede repetirá o cálculo de distâncias até que os pesos sejam bem próximos aos do vetor de entrada. O objetivo deste processo é agrupar dados em diversos grupos próximos ou clusters (COPIN, 2013). Como não definimos previamente a quantidade de classes a serem

identificadas pela rede, os Mapas Auto-Organizáveis se mostraram uma opção apropriada.

4. Resultados

As amostras de dados, como citado anteriormente, provém de questionário MTA-SNAP-IV com os quatro quantificadores. Realizamos dois conjuntos de experimentos (algoritmo de clusterização e rede neural não supervisionada SOM).

O algoritmo de clusterização foi aplicado através do software MatLab (MATLAB, 2016). Como citado anteriormente, é necessário definir previamente o número de agrupamentos. Definimos inicialmente como k=2, e a grande maioria das variáveis foi atraída apenas para um cluster, o que não permitiu uma classificação clara das variáveis. A partir de três clusters (k=3), pudemos classificar diversas variáveis em diversos clusters. Verificamos experimentalmente que com apenas 3 já seria suficiente, pois nosso objetivo era ter as classes: "*possível candidato a TDAH*", "*não é candidato*", "*possível candidato a outro transtorno*".

Na figura 1, percebe-se que o algoritmo produz como resultado uma imagem em 3D, contendo todas as variáveis (círculos menores) e os três clusteres representados pelos círculos maiores (centróides com cores azul, verde e vermelho). As cores das variáveis indicam que elas estão ligadas ao cluster da cor correspondente. Na figura 1, os pequenos pontos da imagem representam todos os indivíduos avaliados. Pode-se observar uma grande proximidade de um indivíduo ao cluster verde.

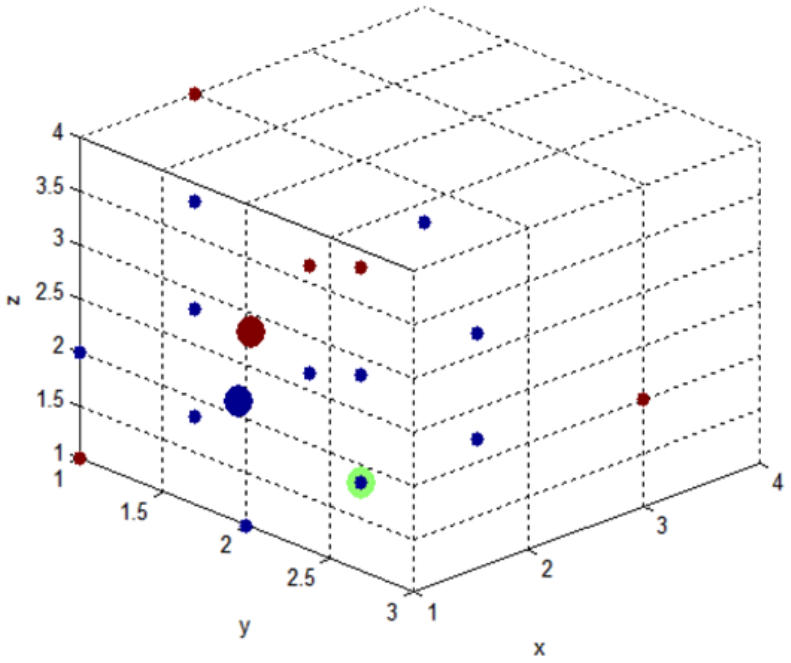


Figura 1. Agrupamento em três clusters (verde, azul e vermelho) do algoritmo *k-means*

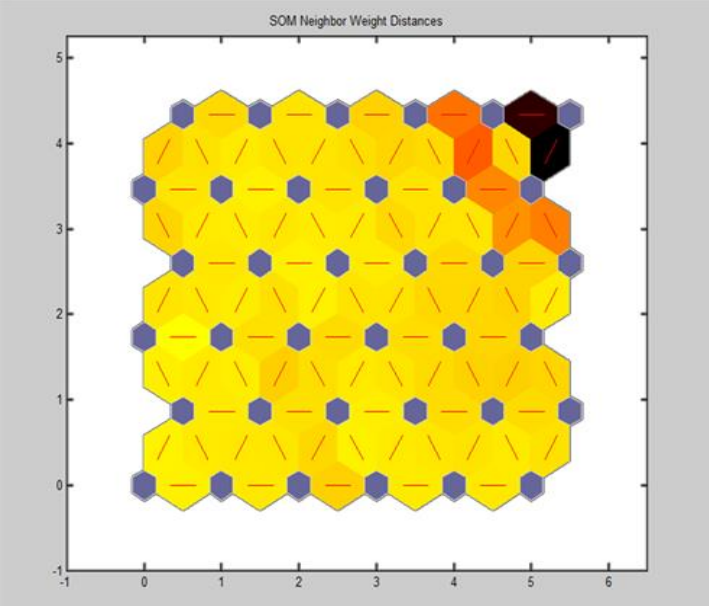


Figura 2. Mapa de vizinhança das variáveis

A rede neural utilizada neste experimento do tipo não supervisionada (Mapas Auto Organizáveis-SOM) foi processada através do software MatLab (MATLAB, 2016). De acordo com a Figura 2, que ilustra o mapa de vizinhanças (Neighbor Weight Distances) como saída da rede SOM, observa-se que foi possível separar em classes registros com caracteristicas bem definidas. Percebe-se a presença marcante de três registros bem próximos situados no canto superior direito do mapa (hexágonos roxo). Porém existe apenas um candidato que se apresenta de maneira acentuada destancando-se dos demais. Este candidato (hexágono) encontra-se situado na extremidade do canto superior direito da Figura 2. Dessa forma, podemos inferir que é possível que estes três indivíduos sejam fortes candidatos em risco de TDAH.

Nestes experimentos estamos interessados em identificar possíveis candidatos em risco de TDAH. Percebe-se que apenas um indivíduo foi identificado como muito próximo a um cluster (Figura 1, círculo verde maior). Na figura 2, o mesmo indivíduo foi classificado pela rede SOM como um forte candidato em risco de TDAH.

Após inúmeras execuções do *k-means* (n > 20), o algoritmo permaneceu exibindo o mesmo resultado. A rede SOM foi executada com 200 épocas, também apresentando o mesmo resultado. Constatou-se que o mesmo o candidato destacou-se nos diferentes algoritmos. Estes resultados permitem inferir que se trata de um forte candidato ao risco de TDAH. Porém, isso só poderá ser confirmado após o diagnóstico clinico do especialista.

5. Discussão

Os resultados obtidos através de nossos experimentos podem ser considerados satisfatórios uma vez que os dois algoritmos convergiram após diversas execuções. Através do algoritmo *k-means* percebemos que um dos indivíduos se destacou sendo o único de uma das tres classes; pudemos concluir que este é um candidato forte ao risco

de TDAH. A rede não supervisionada apontou uma variável como sendo a mais destacada, como visto na figura 2. Assim, percebemos que os resultados dos dois algoritmos apontam para a existência de um mesmo candidato ao risco de TDAH.

Neste trabalho nenhum dos entrevistados passou *a priori* pelo processo de diagnóstico clínico com os especialistas. Portanto, será necessário que os possíveis candidatos apontados pelos experimentos sejam avaliados clinicamente.

Ressaltamos algumas dificuldades encontradas na realização desta pesquisa. Como citado anteriormente, não foram encontrados padrões de individuos diagnosticados, assim como os entrevistados não possuem diagnóstico confirmado de TDAH feito por especialista. Na busca por padrões, foi necessária uma análise mais detalhada dos resultados algoritmicos.

6. Conclusões

O TDAH pode atingir até 5% de crianças em idade escolar e ocasionar perdas acadêmicas e sociais. Desta forma, surge a oportunidade de pesquisar estratégias computacionais voltadas para esse público. Este trabalho investigou o uso de técnicas de Inteligência Artificial baseadas em algoritmo de clusterização *k-means* e em rede neural não-supervisionada do tipo SOM como uma forma de indicar indivíduos em risco de TDAH, auxiliando seu diagnóstico precoce.

Esse trabalho explorou por meio de dois experimentos um dataset composto de dados obtidos de pais respondentes. Os resultados obtidos se mostraram consistentes pois um indivíduo foi indicado como possível candidato em risco de TDAH pelos dois algoritmos avaliados, mesmo após diversas execuções (n>20).

Como trabalhos futuros, pretendemos desenvolver um sistema que incorpore as técnicas de Inteligência Artificial para processar dados de novos possíveis candidatos a partir de padrões de diagnosticados clinicamente. O sistema poderá ser utilizado no ambiente escolar pela equipe pedagógica para auxiliar na detecção precoce de crianças e adolescentes em risco de TDAH.

Referências

APA DSM IV TR – Manual Diagnóstico e Estatístico de Transtornos Mentais, 4ª ed. revisada. Porto Alegre: ArtMed, 2002.

BRASIL. Lei de Diretrizes e Bases da Educação Nacional n. 9394/96, artigo 58, de 20 de dezembro de 1996.

Carvalho, L. A. V. (2005) Datamining – A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. Rio de Janeiro: Editora Ciência Moderna Ltda.

COPPIN, Ben. Inteligência Artificial. Tradução e revisão técnica Jorge Duarte Pires Valério. – [Reimpr.] – Rio de Janeiro: LTC, 2013.

COSTA, M. Uma estratégia computacional na detecção de dislexia. UFRJ, 2011.

COSTA et al. Desafios e Oportunidades em Neurociência Computacional na Educação Brasileira. 3º Seminário Grandes Desafios da Computação. SBC, 2014. Disponível em < https://goo.gl/tCBbh8>. Acesso em 20 de Jun 2016.

Mattos et al. Apresentação de uma versão em Português para uso no Brasil do instrumento MTA-SNAP-IV de avaliação de sintomas de Transtorno de deficit de atenção/hiperatividade e sintomas de transtorno desafiador e de oposição. Revista de Psiquiatria RS, 2006.

MIRANDA, C. T. ; SANTOS JUNIOR, G. ; PINHEIRO, N. A. M. ; STADLER, Rita de Cassia L. . Questionário SNAP-IV: a utilização de um instrumento para identificar alunos hiperativos. In: VIII ENPEC - I CIEC, 2011, Campinas. VIII ENPEC - I CIEC, 2011.

Rohde, L. A., & Ketzer C. R. (1997). Transtorno de déficit de atenção e hiperatividade. In N. Fichtner (Org.),Transtornos mentais da infância e adolescência (pp. 232-243). Porto Alegre: Artmed.

Wazlawick, Raul Sidnei. Metodologia de pesquisa para ciência da computação. Campus, 2ª ed.,2014.

KANUNGO, T. et al., 2002, An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, nº 7, July.

JAIN, A. K., MURTY, M. N. e FLYNN, P., 1999, Data clustering: a review. ACM Computing Surveys 3 (31): 264–323.

MatLab – The Language of Technical Computing. R2016a, MatWorks. Mais informações em www.mathworks.com – Acesso em 2016.

Aplicação de Algoritmos de Árvore de Decisão na Previsão da Evasão Escolar: um estudo no campus Lagarto do IFS

Marília dos Anjos Santos, Rodrigo Fontes Cruz, Lauro Barreto Fontes, Gilson Pereira dos Santos Júnior, Glauco Luiz Rezende de Carvalho

[1]Coordenadoria do Bacharelado de Sistemas de Informação – Campus Lagarto
Instituto Federal de Educação, Ciência e Tecnologia de Sergipe (IFS)
Estrada da Barragem, nº 425, Povoado Carro Quebrado. – Lagarto, SE – Brasil

marilia.annjos@gmail.com, {lauro.barreto, gilson.pereira}@ifs.edu.br

Abstract. *The school evasion rate is one of the chronic problem of education. The integrated course of IFS Computer Networks, campus Lagarto, suffers from evasion of 18.9% of freshment, threatening, thus, the fulfillment of the goals and commitments of greement (TAM) established with the government. To define measures to reduce this value is necessary to trace the profile of the student who evaded. Therefore, we used the J48 decision tree algorithm, available in Weka tool, to analyze the academic and personal data of 210 students of the course. In the present study, we concluded that 81% of evasions occur in the first two years and reaches young people of 17 years or more, with low performance at computer programming,physicis, english or computer networks.*

Resumo. *A evasão escolar é um dos problema crônico da educação. O curso integrado de Redes de Computadores do IFS, Campus Lagarto, sofre com a evasão de 18,9% dos ingressantes, ameaçando, assim, o cumprimento do Termo de Acordo de Metas e Compromissos (TAM) estabelecido com o Governo. Para definir medidas que reduzam este valor é necessário traçar o perfil do aluno que evade. Com este objetivo, utilizou-se o algoritmo de árvore de decisão J48, disponível na ferramenta Weka, para analisar os dados acadêmicos e pessoais de 210 alunos do curso. Com o estudo, concluiu-se que 81% das evasões ocorrem nos dois primeiros anos e atinge jovens de 17 anos ou mais, com baixo rendimento em programação, física, línguas ou redes de computadores.*

1. INTRODUÇÃO

A evasão escolar é um problema crônico da educação brasileira que atinge tanto o setor público quanto o privado, causando prejuízos acadêmicos, financeiros, políticos e sociais aos envolvidos no processo educacional.

Ações governamentais como o Programa REUNI - Reestruturação e Expansão das Universidades Federais [MEC/SESu/REUNI 2007] e o TAM - Termo de Acordo de Metas e Compromissos [MEC/SETEC 2009], estabelecidos, respectivamente, com as Instituições Federais de Ensino Superior (IFES) e os Institutos Federais de Educação, Ciência e Tecnologia (IF), propõem a viabilização de recursos, mediante ao cumprimento de metas, dentre elas, o controle da taxa de evasão. Neste quesito, o TAM estabelece que no mínimo 80% dos discentes que iniciam o curso devem concluí-lo.

Entretanto, o Campus Lagarto do Instituto Federal de Educação, Ciência e Tecnologia de Sergipe (IFS), âmbito deste estudo, vêm enfrentando problemas para controlar a taxa de evasão nos cursos técnicos de nível médio, na modalidade integrado. Atualmente, os cursos de edificações, eletromecânica e redes de computadores apresentam as taxas de 9,32%, 12,18% e 18,09%, respectivamente [IFS 2015].

Assim, diante dos valores apresentados, o curso de Redes de Computadores está no limiar para o não cumprimento do TAM. Portanto, é fundamental traçar o perfil do discente que evadiu, a fim de identificar quais foram as causas e os motivos da evasão e, em seguida, definir e aplicar medidas para minimizar tal prática.

De acordo com a literatura [Manhães et al. 2012, Marquez-Vera et al. 2011, Veitch 2004], a Mineração de Dados (MD) apresenta-se como uma ferramenta capaz de detectar comportamentos ligados à evasão escolar, uma vez que ela extrai informações valiosas, a partir de dados, para descrever o perfil evasivo do discente.

Segundo [Veitch 2004], dentre as técnicas de MD mais utilizadas com tal propósito, a Árvore de Decisão destaca-se por ser um mecanismo eficiente, de baixo tempo de processamento, para criação de classificadores a partir da exploração dos dados. Uma das vantagens desta técnica é que os resultados são representações simbólicas, em formato de árvore, facilitando a compreensão e interpretação por humanos.

Diante do escopo, o objetivo deste trabalho é descobrir o perfil do discente propenso a evadir do curso técnico integrado ao nível médio de Redes de Computadores(IRC) do IFS, campus Lagarto, por meio da utilização de algoritmos de árvore de decisão.

2. TRABALHOS RELACIONADOS

2.1. Previsão da Evasão Escolar com Árvore de Decisão

A evasão escolar é bastante discutida na literatura e os autores normalmente evidenciam a dicotomia entre fatores intrínsecos e extrínsecos à escola como causadores do abandono escolar. Uma das formas de investigar os fatores que afetam a evasão é utilizar técnicas de mineração de dados.

Neste contexto, o estudo desenvolvido por [Veitch 2004] investigou variáveis relacionadas à evasão escolar no ensino médio por meio da técnica de árvore de decisão CHAID (CHi-quadrado Automatic Interação Detection). Neste, o autor observa que o desempenho acadêmico está fortemente relacionado com o abandono escolar e que a técnica utilizada conseguiu prever quais alunos tendem a sair da escola.

Com o mesmo intento, [Marquez-Vera et al. 2011] estuda comparativamente 10 classificadores a fim de verificar qual apresenta melhor acurácia na previsão de alunos propensos ao fracasso escolar. Eles concluíram que a árvore de decisão ADtree apresenta melhor precisão. Além disso, eles observaram que alunos maiores de 15 anos, cursando o período noturno, que possuem mais de um irmão ou irmã, que apresentam deficiências em materiais de Física, de Ciências Humanas, de Matemática e de Inglês e que possuem baixo nível de motivação para os estudos, tendem a abandonar a escola antes da conclusão do curso.

Já em 2012, [Manhães et al. 2012] compararam 6 algoritmos classificadores e verificaram uma acurácia na previsão variou entre 70% e 86%, sendo os melhores resulta-

dos alcançados pelo Multilayer Perceptron (85%), Support Vector machine (86%) e o J48 (83%). A amostra dos dados era composta por 7304 alunos do ensino superior, classificados em: aluno com curso ou matrícula cancelada, aluno com matrícula ativa ou curso em andamento ou, ainda, aluno com matricula concluída. Ao final do estudo, observou-se que alunos com perfil de evasão possuem: ao menos uma disciplina reprovada por falta e média no 1º período, ou no mínimo uma reprovação no 2º período por média, ou redução no número de disciplinas cursadas e aprovadas ao longo do curso, ou as médias de aprovação no 1º período são inferiores aos dos demais alunos, ou ainda, possuem, ao final do 2º período, média similar ou inferior à média geral dos alunos com matrículas canceladas

Assim, diante dos trabalhos supracitados, percebe-se que os algoritmos para geração de árvores de decisão alcançam bons resultados na previsão da evasão escolar. Destaca-se ainda que dados como idade, gênero, etnia, estado civil, renda familiar, atividade profissional, escola de origem, nível de leitura e escrita, conhecimento da matemática, matriz curricular, total de faltas, média geral e por disciplina, total de disciplinas aprovadas por média e por falta, bem como, as reprovadas são fundamentais para traçar o perfil do discente que evade.

3. MATERIAIS E MÉTODOS

3.1. Coleta de dados

O curso integrado de Redes de Computadores (IRC) do IFS, Campus Lagarto, possui entrada anual e sua escolha justifica-se pelo alto índices de evasão, uma vez que, atualmente atinge cerca de 18,9% dos ingressos, de acordo com [IFS 2015].

Os dados coletados para análise da previsão da evasão originam do curso IRC e compreende o período de 4 anos, de 2011/01 até 2014/2. Estes dados foram extraídos do QAcadêmico, sistema de gestão e controle acadêmico integrado, no qual se concentra toda vida acadêmica do corpo discente do IFS.

No estudo foram selecionadas informações pessoais (idade, sexo, etnia, grau de instrução do pai e da mãe), bem como, acadêmicas (ano de ingresso, total de períodos – aprovados, aprovados com dependência, reprovados, reprovados por média ou reprovados por falta –, tempo que gastou para concluir o curso e rendimentos da turma nas disciplinas.)

3.2. Preparação dos Dados

Durante o pré-processamento foram eliminadas as colunas com informações pessoais e acadêmicas que não faziam parte do estudo, conforme definido em Seção 3.1, além de colunas em branco ou com dados duplicados. Após este processo, a planilha resultou em 210 registros de alunos do IRC, sendo 134 registros (63,8%) CURSANDO, 56 registros (26,66%) representando a classe de EVASÃO e apenas 20 registros (9,52%) CONCLUÍDO.

O considerável desbalanceamento entre as classes (cursando, evasão e concluído) na base da dados poderia influenciar negativamente na previsão da árvore de decisão, conforme explicado por [Fugimoto et al. 2009] apud [Batista 2003] quando afirmam que ”classificadores induzidos a partir de classes balanceadas artificialmente, alocando 50%

dos exemplos de treinamento para a classe minoritária, geralmente, apresentam resultados melhores do que aqueles com distribuição natural das classes.”.

Portanto, seguiu-se a metodologia utilizada por [Fugimoto et al. 2009], na qual os dados foram divididos em 10 conjuntos de treinamentos (T1, T2, ..., T10), cada um contendo todos os registros da classe minoritária, mais os dados escolhidos de forma aleatória das demais classes, de forma que todos os conjuntos possuam uma distribuição com representatividade semelhante de classes, conforme Figura 1.

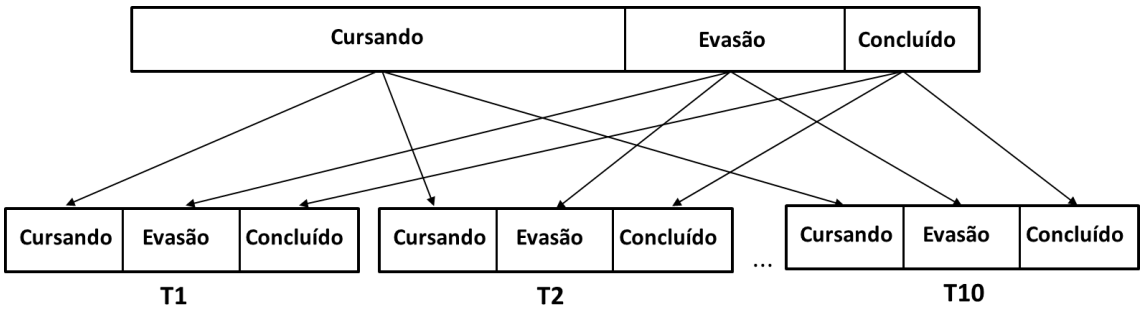


Figura 1. Processo de montagem do conjunto de treinamento inspirado em [Fugimoto et al. 2009].

3.3. Mineração dos Dados

A ferramenta utilizada para auxiliar na mineração foi o Weka. Esta aplicação é totalmente gratuita, código aberto, de fácil utilização e disponibiliza 12 (doze) algoritmos de árvore de decisão implementados, além de possibilitar a inserção de novos.

Inicialmente, realizou-se uma pesquisa experimental com todos os algoritmos de árvore de decisão disponibilizado pelo Weka para selecionar qual deles melhor prévia a evasão escolar, diante dos dados coletados na referida instituição de ensino. Nesta estudo, o algoritmo foi escolhido mediante a taxa de acurácia e medida de kappa. A Tabela 1 ilustra o resultado deste experimento.

Uma vez escolhido o algoritmo, este foi executado com cada uma das 10 bases de treinamento balanceadas durante a preparação dos dados. Obteve-se, portanto, um total de 10 árvores, sendo que cada uma descreve o perfil evasivo do conjunto de dados analisado. Em seguida, foram selecionadas as 3 (três) árvores que apresentaram o maior valor de medida Kappa, maior acurácia, menor número de folhas e menor tamanho da árvore.

4. RESULTADOS E DISCUSSÕES

4.1. Seleção do Algoritmo de Árvore de Decisão

O resultado da execução dos algoritmos de árvore de decisão disponibilizados pelo Weka corroborou com as informações encontradas em [Manhães et al. 2012], visto que o J48 foi o mais consistente, atingindo 0,9065 de medida Kappa e 95% de acurácia, conforme apresentado na Tabela 1. Além do J48, os algoritmos LadTree e Random Forest apresentaram resultados expressivos.

Embora o J48 tenha sido selecionado para traçar o perfil evasivo dos estudantes, neste estudo, os algoritmos LadTree e Random Forest também apresentaram resultados

Tabela 1. Seleção do algoritmo de árvore de decisão.

Algoritmo	Kappa	Acurácia	Algoritmo	Kappa	Acurácia
BFtree	0,8592	93%	LMT	0,8142	88%
DecisionStump	0,6163	83%	NbTree	0,8583	93%
FT	0,7970	89%	RandomForest	0,8760	94%
J48	0,9065	95%	RadomTree	0,6587	83%
J48Graft	0,8472	94%	RepTree	0,8541	93%
LadTree	0,8788	95%	SimpleCart	0,8359	92%

expressivos. Por outro lado, os algoritmos Decision Stump e Random Tree se demonstraram pouco interessante para esta tarefa, principalmente, se analisado o valor da medida Kappa.

4.2. Extração do Perfil de Evasão com o Algoritmo J48

Os resultados obtidos com a execução do algoritmo J48 em cada uma das 10 (dez) bases estão apresentados na Tabela 2.

Tabela 2. Árvores de decisão geradas pelo algoritmo J48.

Árvore	Kappa	Acurácia	Nós	Tamanho	Variáveis Decisórias
A1	0,5549	71%	6	11	Idade, Química, Física
A2	0,5775	72%	5	9	Idade, Português, Microinformática
A3	0,5220	70%	5	7	Idade, Matemática
A4	0,5166	67%	5	9	Idade, Física, Química
A5	0,6161	74%	6	11	Idade, Português
A6	0,5932	72%	3	5	Idade, Ling. de Programação
A7	0,4948	66%	3	5	Idade, Ling. de Programação
A8	0,5547	70%	5	9	Idade, Ling. Programação, português
A9	0,7367	82%	4	7	Idade, Inglês, Ling. de Programação
A10	0,5935	73%	5	9	Idade, Redes, Física

Conforme ilustrado na Tabela 2, as árvores A5, A9 e A10 apresentam os melhores resultados, se considerada a medida Kappa, a acurácia, o número de nós e o tamanho da árvore. Estas árvores podem ser visualizadas nas Figura 2, 3 e 4, respectivamente.

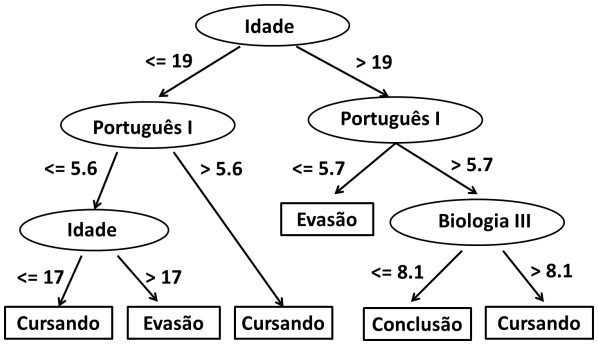


Figura 2. Árvore de Decisão A5 gerada pelo algoritmo J48.

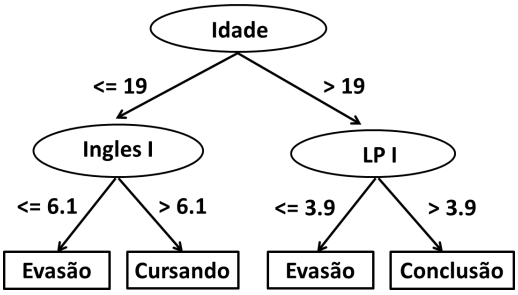


Figura 3. Árvore de Decisão A9 gerada pelo algoritmo J48.

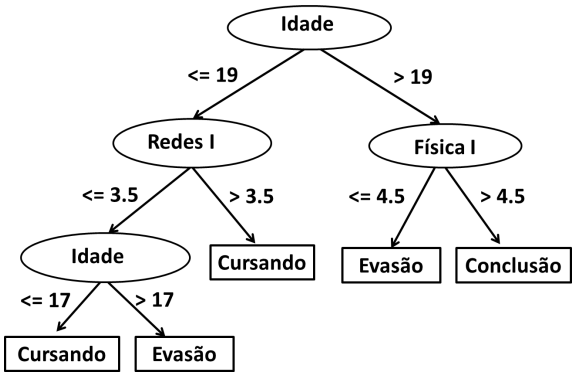


Figura 4. Árvore de Decisão A10 gerada pelo algoritmo J48.

É importante ressaltar que, embora a árvore A10 se destaque como uma das melhores no resultado geral, o seu valor de Kappa apresenta uma concordância apenas moderada. Já a árvore A5 é um pouco melhor, pois possui um nível de concordância substancial, assim como, a árvore A9. Esta última, entretanto, apresenta o melhor resultado de Kappa e de acurária e o total de nós, bem como, o tamanho da árvore é menor do que A5 e A10.

Analisando as árvores foi possível perceber incidência de característica entre as variáveis decisórias como, por exemplo, a idade que está presente em cada uma das 10 árvores geradas. Além disso, as disciplina de física, química, português e linguagem de programação estão presentes em mais de uma árvore.

A partir das árvores é possível inferir as seguintes regras:

$$(Idade > 17) \wedge (PortuguesI \leq 5,7) \rightarrow Evasao \quad (1)$$

$$(Idade \leq 19) \wedge (InglesI \leq 6,1) \rightarrow Evasao \quad (2)$$
$$(Idade > 19) \wedge (LingdeProgramacaoI \leq 3,9) \rightarrow Evasao$$

$$(Idade > 17) \wedge (Idade \leq 19) \wedge (RedesdeComputadores \leq 3,5) \rightarrow Evasao \quad (3)$$
$$(Idade > 19) \wedge (FisicaI \leq 4,5) \rightarrow Evasao$$

As regras 1, 2 e 3 foram extraídas das árvores de decisão A5, A9 e A10, respectivamente. Elas, por sua vez, afirmam que alunos com baixo rendimento nas disciplinas de

linguagem de programação, português, física, inglês e redes de computadores do primeiro ano tem grande probabilidade de evadir. As chances de abandono escolar aumentam para alunos mais velhos, com idade equivalente ou superior a 17 anos.

Vale frisar que as disciplinas supracitadas pertencem à grade do primeiro ano do curso. Tal fato, destaca que este ano define o futuro do estudante, conforme apresentado na Figura 5. Observa-se que a evasão acontece ainda nos primeiros anos, sendo 34% no primeiro e 43% no segundo ano cursado. Destes 34% que evadem no primeiro ano de curso, apenas 15% chegam a finalizar as disciplinas, os demais saem antes mesmo de concluir as primeiras avaliações.

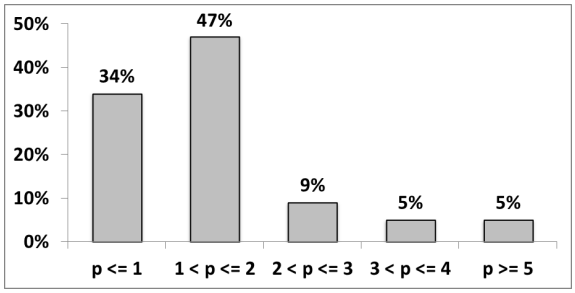


Figura 5. Taxa de evasão no IRC ao longo do curso.

5. CONSIDERAÇÕES FINAIS

A literatura comprovou que o problema da evasão é muito complexo, influenciado por inúmeras variáveis, que afeta o mundo todo e, mesmo assim, ainda não se conhece uma solução. Entretanto, a técnica de árvore de decisão da mineração de dados se apresenta como uma forte aliada na identificação das causas e descrição do perfil do discente que pretende evadir.

A aplicação do algoritmo de árvore de decisão, J48, nos dados pessoais e acadêmicos dos discentes do curso integrado de Redes de Computadores do IFS, Campus lagarto, constatou que é fundamental realizar medidas para reduzir a evasão nos 2(dois) primeiros anos do curso, principalmente com os alunos, de 17 anos ou mais, que obtiveram rendimento inferior nas disciplinas de programação, física, línguas e redes de computadores no primeiro ano.

Esta informação é valiosa para a direção e a gerência de ensino do referido Campus, uma vez que aponta: (i) qual o perfil do discente que necessita de acompanhamento da assessoria pedagógica e que deve participar nas ações de combate a evasão; (ii) bem como as disciplinas nas quais os docentes devem monitorar o desempenho dos alunos, por serem indicadores de um provável o abandono escolar.

Futuramente, pretende-se utilizar o J48 para predizer o perfil de evasão dos demais cursos da instituição, bem como, ampliar o estudo experimentando os algoritmos LadTree e Random Forest, uma vez que eles também apresentaram bons resultados na avaliação preliminar para escolha do algoritmo de árvore de decisão.

Referências

Batista, G. (2003). Pré-processamento de dados em aprendizado de máquina supervisionado. *Instituto de Ciências Matemáticas e de Computação, ICMC. São Carlos, SP*.

Fugimoto, P. M., Sales, L. D. F., Júnior, G. A. P., Passos, A. D. C., Alves, D., and Baranauskas, J. A. (2009). Análise comparativa entre árvores de decisão e triss na predição de sobrevida de pacientes traumatizados. In *IV Congresso da Academia Trinacional de Ciências*.

IFS (2015). Taxa de evasão dos cursos técnicos de nível médio, na modalidade integrado, do campus lagarto do instituto federal de educação, ciência e tecnologia de sergipe.

Manhães, L. M. B., Cruz, S., Costa, R. J. M., Zavaleta, J., and Zimbrão, G. (2012). Identificaçao dos fatores que influenciam a evasao em cursos de graduaçao através de sistemas baseados em mineraçao de dados: Uma abordagem quantitativa. *Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo*.

Marquez-Vera, C., Romero, C., and Ventura, S. (2011). Predicting school failure using data mining. In *EDM*, pages 271–276.

MEC/SESu/REUNI (2007). Reestruturação e expansão das universidades federais.

MEC/SETEC (2009). Termo de acordo de metas e compromissos.

Veitch, W. R. (2004). Identifying characteristics of high school dropouts: Data mining with a decision tree model. *Online Submission*.

Parametrização de Operadores Genéticos na Resolução do Problema de Escalonamento de Horários

Thiago dos Santos[1], J. Francisco S. Neto[1], Gilson P. Santos Júnior[1], Lauro B. Fontes[1], Thiers G. R. Sousa[1]

[1]Coordenadoria de Informática – Instituto Federal de Sergipe (IFS)
Rodovia Lourival Batista – S/N – Povoado Carro Quebrado – 49.400-000 – Lagarto – SE – Brazil

programadorthi@gmail.com, jfrancisco.neto@outlook.com, gilson.universidade@gmail.com

Abstract. *The creation of the timetables requires time, effort and a lot of patiance, in addition to be prone to errors. So, this problem attracted attention from scholars and many approaches were proposed to solve the problem in a automatic way, using techniques like Genetic Algorithm. When using this technique is essential to choose the parameters of the genetic algorithm correctly, otherwise this could lead to hurt the performance of the algorithm. In this sense, this paper has investigated the convergence of the traditional genetic algorithm implementation, to solve the timetabling problem, using adjustments in the parameters of the genetic selection, crossover and mutation.*

Resumo. *A criação de grades de horários requer tempo, esforço e muita paciência, além de ser propenso a erros. Por isso, esse tipo de problema atraiu a atenção dos estudiosos e diversas abordagens foram propostas para resolvê-lo de forma automática, usando técnicas como o Algoritmo Genético. No uso desta técnica é fundamental escolher adequadamente os parâmetros do algoritmo genético, visto que uma seleção indevida pode prejudicar o desempenho do algoritmo. Neste sentido, o presente trabalho investigou a convergência de um algoritmo genético tradicional, para resolver o problema de escalonamento de grade de horário, mediante ajustes na parametrização dos operadores genéticos de seleção, cruzamento e mutação.*

1. Introdução

Criação de grade de horários é um problema comum em instituições de ensino, sejam elas de nível básico, médio ou superior. Sempre que há a necessidade de se criar tal grade um problema é criado, já que interesses pessoais entram em jogo. Segundo Schaerf (1999), a criação de grades de horários demanda muitos recursos, tempo e esforço, embora nem sempre alcance um resultado satisfatório como, por exemplo, a grade pode impossibilitar um estudante de cursar uma ou mais matérias do seu interesse; o professor pode ser escalado em dias e horários que está indisponível ou alocado com uma jornada excessiva ou indicado para ministrar uma disciplina na qual não possui capacidade técnica; a turma pode ser indicada para uma sala de aula ou laboratório sem os recursos necessários ou com a capacidade insuficiente.

De acordo com Almeida (2015), a busca por uma solução ótima, em um tempo aceitável para o problema de escalonamento de grade de horário, é inviável, uma vez que as variáveis envolvidas neste (turma, professores, disciplinas etc.) afetam exponencialmente na sua complexidade. Assim, esse problema é categorizado como NP-Completo, cuja principal característica é a inexistência de pelo menos um algoritmo que retorne uma solução ótima do problema em tempo polinomial.

Outro agravante é a natureza de difícil generalização do problema, uma vez que o conjunto de restrições muda de acordo com o perfil da instituição de ensino (Sousa; Costa; Guimarães, 2002). Este fato explica a quantidade de publicações sobre o tema.

A complexidade do problema e a impossibilidade de generalização da solução fazem com que o tema seja amplamente estudado pela comunidade acadêmica. Existe, atualmente, uma literatura extensa de trabalhos envolvendo a resolução do problema de escalonamento de horários utilizando técnicas da Inteligência Artificial, em especial, as evolucionárias como o algoritmo genético (AG). Assim, é possível encontrar na literatura estudos que aplicam exclusivamente o algoritmo genético para resolver o problema (Huynh; Pham; Pham, 2012) (Rodriguez, 2014) (Almeida, 2015), bem como trabalhos que combinam o AG com outras técnicas (Phuc; Khang; Nuong, 2011) (Yang; Jat, 2011) (Verma; Garg; Bisht, 2012).

O Algoritmo Genético (AG) é uma técnica de otimização e busca inspirado no mecanismo de evolução dos seres vivos, que foi introduzido por Holland (1975). O seu funcionamento consiste em evoluir os indivíduos da população por meio de operadores genéticos (seleção, cruzamento e mutação) até que a solução ótima seja encontrada ou as condições de parada sejam atingidas. Assim, o desempenho desta técnica depende da população inicial, que normalmente é criada de forma aleatória, e da parametrização dos operadores genéticos. Pinho, Montevechi e Marins (2009) ressaltam que "a escolha dos parâmetros do algoritmo genético não deve ser feita de forma genérica. Uma escolha mal feita pode prejudicar o desempenho do algoritmo na obtenção dos resultados esperados".

Nesse sentido, o presente trabalho tem como objetivo investigar a convergência de um algoritmo genético tradicional, para resolver o problema de escalonamento de grade de horário, mediante ajustes na parametrização dos operadores genéticos de seleção, cruzamento e mutação. Desse modo, será possível entender como os operadores genéticos interagem entre si e influenciam na convergência do AG, impactando no tempo de resposta e na qualidade da solução.

2. Trabalhos Relacionados

Pinho, Montevechi e Marins (2009) estudaram a influência do tamanho da população (50 e 200), taxa de cruzamento (0,1 e 0,8) e a taxa de mutação (0,1 e 0,3) no desempenho do algoritmo genético por meio de um experimento 2k Fatorial Completo com 3 replicações, 3 fatores com 2 níveis cada. Ao final do estudo, os autores concluíram que apenas a interação entre os fatores AC (Tamanho da população * taxa de mutação) possuem efeito significativo, considerando um nível de significância de 10% e que o "Tamanho da População" e "Taxa de Cruzamento" possuem forte efeito positivo sobre o valor máximo da função de aptidão.

Huynh, Pham e Pham (2012) utilizaram um AG tradicional para resolver o problema de escalonamento de horários de apresentação de dissertações de mestrado na Universidade de Ciência e Tecnologia de Hanói. Os autores combinaram diferentes quantidades de ponto de corte com a mutação de tamanho fixo, que inverte uma quantidade fixa de genes, e a mutação randômica, que sorteia a quantidade de genes aleatórios a serem invertidos. O algoritmo foi testado com dados reais e apresentou soluções viáveis, embora o desempenho não tenha sido comparado.

Rodriguez (2014) propõe uma abordagem utilizando algoritmos genéticos para resolver o problema de escalonamento de horários no Instituto Tecnológico de Zitácuaro no México. Ele utilizou uma codificação tradicional, ou seja, sem efetuar combinações ou customizações. A estrutura do cromossomo era um vetor contendo o professor, a sala, tempo inicial e tempo final. Segundo o autor, os melhores resultados foram obtidos estabelecendo as taxas de 80% e de 30% para mutação e cruzamento, respectivamente. Esta configuração contraria o senso comum encontrado na literatura, uma vez que altas taxas de mutação podem prejudicar a convergência do algoritmo pela perda das características (genes) herdadas durante o cruzamento.

3. Solução Proposta

Para o presente estudo, um algoritmo genético tradicional com elitismo foi codificado na linguagem Java, conforme especificado no Algoritmo 1. O elitismo foi adicionado para garantir que os melhores indivíduos da solução corrente estejam presentes na próxima geração, garantindo, deste modo, as características que os tornam boas soluções.

O cromossomo foi estruturado como vetor que representa uma grade de horário, sendo que cada gene possui um objeto com 3 (três) dados: dia e horário, sala de aula e professor. A fusão do dia com o horário justifica-se pelo objetivo de reduzir a complexidade da solução, mediante diminuição de variáveis.

No Algoritmo 1, a população inicial é construída aleatoriamente e a função de aptidão é inversamente proporcional a quantidade de restrições violadas pelo indivíduo, assim, quanto mais violações a solução tiver, menos apta ela será. O algoritmo é executado enquanto a quantidade máxima de gerações e a aptidão mínima do indivíduo buscado não forem atingidas (linha 3). A seleção dos pais (linha 4) e os operadores genéticos de cruzamento e mutação (linhas 5 e 6, respectivamente) foram definidos para evolução da população. Ao final da execução, o algoritmo retorna o indivíduo melhor adaptado, ou seja, a grade de horário com menor quantidade de restrições violadas.

Os métodos de seleção codificados foram: por torneio, ranking e roleta viciada. A seleção por torneio cria uma população temporária, com um tamanho previamente especificado, e seleciona o melhor indivíduo dessa população. Na seleção por ranking os indivíduos são organizados em ordem decrescente de aptidão e, em seguida, seleciona-se o indivíduo posicionado no topo deste ranking. Já na seleção por roleta, os indivíduos são selecionados através de uma probabilidade que é diretamente proporcional a sua aptidão.

Com relação ao cruzamento, este pode ser de um ponto ou uniforme, sendo que ambos foram codificados para geração de um único filho. O cruzamento de um ponto

consiste em um corte aleatório nos pais, selecionados através de um dos métodos de seleção, seguido da recombinação genética para gerar o filho, contendo a parte inicial do primeiro pai e a parte final do segundo pai. Já no cruzamento uniforme, dois indivíduos, previamente selecionados através de um dos métodos de seleção, são recombinados por meio de um sorteio para decidir de qual indivíduo o gene será herdado. O sorteio é realizado gerando um número entre 0 e 1 que se o número obtido for menor que 0.5, pega-se o gene do primeiro pai, caso contrário, pega o gene do segundo pai. Esse sorteio é realizado para cada novo gene do indivíduo filho.

Algoritmo 1 Algoritmo genético para escalonamento de horários

```
1:  input: Os dados do problema D;
2:  p ← ag.inicializarPopulacao(input, tamanhoDaPopulacao);
3:  while (geracao < qtdMaximoDeGeracoes and ag.aptidao(p) < limite) do
4:      pais ← ag.selecionarPais(p, metodoSelecao);
5:      filhos ← ag.aplicarCruzamento(pais, metodoCruzamento, taxaDeCruzamento);
6:      filhos ← ag.aplicarMutacao(filhos, metodoMutacao, taxaDeMutacao);
7:      p ← ag.atualizarPopulacao(p, filhos);
8:      geracao ← geracao + 1;
9:  end while
10: return mais apto;
```

Foram codificadas ainda as mutações uniforme e por swap. Na mutação uniforme, um gene k do cromossomo é escolhido aleatoriamente e substituído por outro gene sorteado dentro dos valores possíveis para o parâmetro k. Enquanto a mutação por swap, também conhecida como mutação por troca de argumentos, ocorre a troca de genes de um mesmo cromossomo.

4. Material e Métodos

Para realização deste trabalho foi executada uma pesquisa de natureza aplicada e origem quantitativa, com o objetivo de investigar o comportamento dos parâmetros de configuração dos operadores genéticos. Quanto ao procedimento foi executado um experimento fatorial completo com 4 fatores e 50 replicações, em que a quantidade de níveis variou a depender do fator, conforme apresentado Tabela 1.

Para avaliação foi criado um dataset sintético contendo 13 salas, 4 professores, 6 matérias e 10 turmas. A distribuição de alunos em cada turma segue as informações da Tabela 2.

Os experimentos foram conduzidos em notebook i5 com 6GB de memória RAM. Cada uma das 60 combinações entre fatores e níveis foi executada 50 vezes. Os resultados obtidos em cada execução eram armazenados automaticamente em um

arquivo texto para posterior análise dos dados. As informações armazenadas foram: configuração do teste, indicando a combinação de fatores e níveis; sequência da replicação; aptidão da solução encontrada e; total de gerações executadas.

Tabela 1. Fatores e níveis do experimento. IFS, 2016

Fator	Descrição	Níveis
A	Método de Seleção	Ranking (Ra), Torneio (To) e Roleta Viciada (Ro)
B	Cruzamento	Cruzamento de um ponto (Po) e Uniforme (U)
C	Mutação	Swap (Sw) e Uniforme (Mu)
D	Taxa de (Cruzamento, Mutação)*	(50%, 50%), (60%, 40%), (70%, 30%), (80%, 20%) e (90%, 10%)

* A taxa de cruzamento e mutação é um par ordenado, no qual o primeiro valor refere-se a taxa de cruzamento da população, enquanto o segundo indica a taxa de mutação.

Além disso, durante as execuções, o tamanho da população foi fixado em 100 indivíduos e a evolução foi limitada a 1000 gerações, buscando sempre a aptidão ótima e garantindo 2% de elitismo.

Tabela 2. Distribuição dos alunos por turmas criadas no dataset. IFS, 2016.

Turma	1	2	3	4	5	6	7	8	9	10
Quantidade de Alunos	10	30	18	25	20	22	16	18	24	25

5. Resultados e Discussões

Após a execução dos experimentos, os dados foram tabulados para análise exploratória. Para análise da aptidão dos indivíduos foram utilizadas as métricas de valores mínimos (*min*) e máximos (*max*), média da população (μ), desvio-padrão (*s*) e percentual de soluções ótimas encontradas, ou seja, o percentual de soluções que atingiram aptidão igual a 1. Com relação ao número de gerações foi analisada a média da população (μ) e o desvio-padrão (*s*). Na Tabela 3 estão tabulados 10 (dez) melhores resultados, organizados em ordem decrescente, com relação ao percentual de soluções ótimas alcançadas.

Como pode ser analisado na Tabela 3, os 10 (dez) melhores resultados se restringiram a combinação da seleção por torneio ou roleta viciada (Fator A), cruzamento uniforme (Fator B), mutação por swap (Fator C), sendo que a taxa de cruzamento e mutação (Fator D) variou bastante. Observa-se ainda que, neste estudo, a seleção por torneio obteve uma pequena vantagem se comparada à seleção por roleta viciada com relação ao percentual de soluções ótimas encontradas.

Outro ponto a ser destacado é que as combinações com torneio necessitaram, em média, um quantitativo menor de gerações para finalizar a execução do algoritmo. Este fato está diretamente associado: (i) ao percentual de solução ótima encontrada, uma vez

que quando uma solução perfeita é obtida, o algoritmo encerra a execução; (ii) a população inicial gerada aleatoriamente.

É importante ressaltar ainda que a mutação por swap, em conjunto com o cruzamento uniforme, tem uma contribuição muito grande para a taxa de sucesso do algoritmo, já que se pode permutar o método de seleção entre torneio e roleta viciada, sem grandes alterações nos resultados.

Além disso, embora exista um entendimento na literatura que (i) altas taxas de mutação prejudiquem o desempenho do algoritmo, devido à perda das características genéticas das gerações; e (ii) taxas muito pequenas de mutação prejudiquem a convergência do algoritmo, uma vez que não garantem a variabilidade da população, o presente estudo não alcançou resultados conclusivos sobre esta questão. Isso porque, a contribuição da combinação de cruzamento uniforme e mutação por swap não permitiram identificar um impacto significativo da variação na taxa de cruzamento e mutação.

Tabela 3. Relações dos operadores e suas configurações com taxas de sucesso

Fatores				Aptidão				Soluções Ótimas	Gerações	
A	B	C	D	μ	*s*	*max*	*min*		μ	s
To	U	Sw	(0,8, 0,2)	0,90	0,21	1,00	0,33	82%	243,22	379,00
To	U	Sw	(0,5, 0,5)	0,91	0,20	1,00	0,33	82%	250,14	357,64
Ro	U	Sw	(0,7, 0,3)	0,89	0,22	1,00	0,33	80%	375,06	344,05
To	U	Sw	(0,7, 0,3)	0,90	0,21	1,00	0,25	80%	261,12	380,99
To	U	Sw	(0,9, 0,1)	0,87	0,23	1,00	0,33	76%	285,54	411,47
Ro	U	Sw	(0,6, 0,4)	0,86	0,23	1,00	0,33	72%	457,90	350,39
To	U	Sw	(0,6, 0,4)	0,85	0,24	1,00	0,33	70%	379,70	427,03
Ro	U	Sw	(0,8, 0,2)	0,84	0,24	1,00	0,50	68%	439,72	414,26
Ro	U	Sw	(0,5, 0,5)	0,77	0,28	1,00	0,33	58%	602,14	349,25
Ro	U	Sw	(0,9, 0,1)	0,75	0,28	1,00	0,25	54%	541,56	443,66

6. Conclusões

A criação automática de grade de horários utilizando técnicas de Inteligência Artificial, em especial algoritmos evolucionários, ainda é uma área com espaço para inovação, principalmente com soluções híbridas, adaptações ou combinações.

O presente estudo se limita pela utilização de um conjunto de dados (dataset) sintéticos, construído por meio de valores aleatórios. Portanto, possível que estes dados tenham influenciado na dificuldade de observar um impacto significativo na variação da taxa de cruzamento e mutação, pois, mesmo buscando representar dados reais, nosso (dataset) tem pouca informação comparada à uma situação real.

Assim, como trabalhos futuros, pretende-se executar novos experimentos visando focar a investigação no impacto da variação da taxa de cruzamento e mutação de forma isolada. Avaliar o desempenho desta codificação por meio de benchmarks de datasets públicos. Além disso, avaliar o desempenho do algoritmo genético combinado a outras técnicas de otimização, a exemplo das abelhas artificiais, da colônia de abelhas, da busca tabu, da subida pela encosta etc.

7. Referências

Almeida, Maria Weslane de Sousa. "Utilização de algoritmos genéticos para montagem de horários acadêmicos com foco na blocagem de horários". 2015. 158 f. Trabalho de Conclusão (Curso de Graduação em Sistemas de Informação)-Universidade Federal do Rio Grande do Norte. Caicó, 2015.

Holland, John H. "Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence". U Michigan Press, 1975.

Huynh, Thi Thanh Binh; Pham, Quang Dung; Pham, Duy Dat. Genetic algorithm for solving the master thesis timetabling problem with multiple objectives. In: *2012 Conference on Technologies and Applications of Artificial Intelligence.* IEEE, 2012. p. 74-79.

Pinho, Alexandre Ferreira de; Montevechi, José Arnaldo Barra; Marins, Fernando Augusto Silva. Análise da aplicação de projeto de experimentos nos parâmetros dos algoritmos genéticos. *Sistemas & Gestão*, v. 2, n. 3, p. 319-331, 2009.

Phuc, Nguyen Ba; Khang, Nguyen Tan Tran Minh; Nuong, Tran Thi Hue. A New Hybrid GA-Bees Algorithm for a Real-world University Timetabling Problem. In: *2011 International Conference on Intelligent Computation and Bio-Medical Instrumentation*. 2011.

Rodriguez, Noel et al. Solving a Scholar Timetabling Problem Using a Genetic Algorithm-Study Case: Instituto Tecnologico De Zitacuaro. In: *Artificial Intelligence (MICAI), 2014 13th Mexican International Conference on*. IEEE, 2014. p. 197-202.

Schaerf, Andrea. A survey of automated timetabling. Artificial intelligence review, v. 13, n. 2, p. 87-127, 1999.

Souza, Marcone Jamilson Freitas; Costa, F. P.; Guimarães, I. F. Um algoritmo evolutivo híbrido para o problema de programação de horários em escolas. *XXII Encontro Nacional de Engenharia de Produção-ENEGEP*, p. 8, 2002.

Verma, Om Prakash; Garg, Rohan; Bisht, Vikram Singh. Optimal time-table generation by hybridized bacterial foraging and genetic algorithms. In: *Communication Systems and Network Technologies (CSNT), 2012 International Conference on. IEEE*, 2012. p. 919-923.

Yang, Shengxiang; Jat, Sadaf Naseem. Genetic algorithms with guided and local search strategies for university course timetabling. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, v. 41, n. 1, p. 93-106, 2011.

MobIES: Aplicativo Integrado de Serviços para Instituições de Ensino Superior

Laura K. Engelmann[1], Leonardo A. Sápiras[1]

[1]Faculdades Integradas de Taquara (FACCAT)
Taquara – RS – Brasil

{lauraengelmann,sapiras}@faccat.br

Resumo. *Este artigo apresenta os resultados sobre o desenvolvimento de um aplicativo móvel integrado de serviços voltado para Instituições de Ensino Superior, denominado MobIES. A aplicação tem como finalidade conectar instituição e aluno, fornecendo ao público acadêmico serviços institucionais que facilitam o processo de acompanhamento de informações e de atendimento. Para o funcionamento do aplicativo, o próprio sistema de gestão acadêmica das instituições é utilizado como fonte de dados e a integração é feita através de web services.*

Abstract. *This article introduces the results about the development of an integrated mobile application services oriented for higher education institutions, named MobIES. The application aims to connect student and institution, providing the academic public institutional services that facilitate the process of monitoring information and service. For the operation of application, the own academic management system of institutions is used as data source and the integration is done through web service.*

1. Introdução

Em virtude da necessidade crescente das pessoas estarem permanentemente conectadas, o mercado de dispositivos móveis se expandiu. Isso também pode ser justificado pela característica da portabilidade que os dispositivos móveis proporcionam, permitindo que pessoas fiquem informadas de uma maneira facilitada [Nascimento 2013]. Existem diversos tipos de aplicações móveis, para os mais diversos fins. O presente artigo aborda aplicativos móveis voltado para o oferecimento de serviços de instituições de ensino superior para seu público acadêmico.

No estado do Rio Grande do Sul, por exemplo, observa-se que algumas instituições de ensino superior, tais como UCS[1], UFRGS[2] e FACCAT[3], não oferecem aos seus alunos serviços por meio de aplicativos, apesar de algumas operações poderem ser realizadas por meio de sistemas *web*, telefone ou atendimento presencial. Instituições, como a Feevale, possuem aplicações *mobile*, entretanto tais aplicações funcionam somente nestas instituições e não podem ser reutilizadas ou adaptadas para outras instituições de ensino, pois tais aplicativos foram concebidos para atender apenas as regras de negócio dessas IES.

[1] UCS: Universidade de Caxias do Sul.
[2] UFRGS: Universidade Federal do Rio Grande do Sul.
[3] FACCAT: Faculdades Integradas de Taquara.

No mercado existem poucas soluções *mobiles* voltadas para IES, que poderiam ser integradas aos sistemas de gestão acadêmica das instituições e que conseguiriam suprir todas as funcionalidades que as mesmas possuem. Apesar de não terem sido encontradas na literatura evidências que justifiquem a existência de poucos aplicativos comuns a várias IES, entrevistas realizadas com técnicos administrativos de algumas IES demonstraram que isso ocorra em razão das regras de negócios das instituições, que se diferem em alguns aspectos. Um exemplo de uma regra de negócio assim, são as notas ou avaliações existentes em um semestre. Na FACCAT, cada semestre possui avaliações de Grau 1, Grau 2, substituição de nota de um dos graus, e exame final. Já na Univates, também localizada no Rio Grande do Sul, as avaliações em um semestre são divididas em três avaliações (Nota 1, nota 2 e nota 3).

O presente artigo apresenta os resultados de um trabalho de conclusão, no curso de Bacharelado em Sistemas de Informação da FACCAT. O trabalho teve como objetivo o desenvolvimento de uma aplicação móvel híbrida, denominada MobIES, que permita oferecer ao público acadêmico serviços instituições tais como alteração de dados cadastrais, consulta ao histórico escolar, inscrição em cursos de extensão e acesso a informações financeiras. O diferencial desse projeto é que o sistema desenvolvido tem como premissa ser compatível com qualquer instituição de ensino, independente das regras de negócio existentes e desde que os sistemas das IES possam ser adaptados para isso. Os objetivos específicos são: (i) desenvolver um aplicativo contemplando um conjunto limitado de funcionalidades e, (ii) realizar um estudo de caso da ferramenta em uma IES, para verificar se foi possível ao aplicativo gerenciar regras de negócio dessa IES.

Observa-se que existem aplicações móveis EaD (Ensino à Distância) para instituições de ensino que tem como finalidade servir de apoio didático entre acadêmico e professores. No entanto este não é o propósito deste trabalho, que tem como escopo as aplicações móveis voltadas para o fornecimento de serviços institucionais. Durante o desenvolvimento desta pesquisa, foi identificado que a aplicação proposta possui semelhanças à sistemas CRM (*Custumer Relationship Management*), descritas com mais detalhes em [Gomes et al. 2014].

O presente artigo está estruturado da seguinte forma, a Seção 2 apresenta o referencial teórico, que descreve os assuntos abordados e relacionados sobre o tema em questão. A Seção 3 descreve a metodologia utilizada. Já na Seção 4, são apresentados os resultados obtidos e, por fim na Seção 5, as conclusões sobre a pesquisa realizada.

2. Referencial teórico

2.1. Tipos de aplicativos móveis

Apps ou aplicativos móveis, são *softwares* projetados e desenvolvidos para serem executados e acessados através de dispositivos móveis, como celulares, *smartphones* e *tablets*, tendo como possibilidade o acesso de conteúdos *on-line* e *off-line*. Os apps tem como objetivo facilitar o desempenho de atividades práticas do usuário e podem ser divididos em várias categorias, como, aplicativos de entretenimento, educação, música, saúde, dentre outros [Nonnenmacher 2012]. Diferentes plataformas tecnológicas, incluindo sistemas operacionais e plataformas de desenvolvimento dominam o mercado de dispositivos móveis, fornecendo assim diferentes tipos de soluções de desenvolvimento [Martins et al. 2013]. Existem três principais soluções para o

desenvolvimento de aplicações *mobiles*: as web apps, os aplicativos nativos e os aplicativos híbridos que consistem respectivamente em soluções *webs* formatadas para serem acessadas através do *browser* do dispositivo móvel, soluções desenvolvidas para um específico sistema operacional, ou seja, para uma específica plataforma e soluções desenvolvidas com a junção de aplicativos nativos e web apps [Silva et al. 2015].

Os aplicativos móveis *web* são aplicações que não necessitam serem instaladas no dispositivo móvel. O acesso é feito através do navegador *web* (*browser*) do dispositivo, onde deve ser digitado a URL[4] correspondente à aplicação. [Martins et al. 2013] descrevem que os aplicativos móveis *web*, não conseguem acessar os recursos e funcionalidades da plataforma do dispositivo, de *hardware* e *software*, o que restringe funcionalidades e limita recursos muitas vezes essenciais para aplicações móveis.

Os aplicativos nativos são aplicações desenvolvidas para um tipo específico de plataforma. Existem diversas plataformas para o desenvolvimento de aplicativos móveis, onde cada uma exige que os aplicativos nativos sejam desenvolvidos utilizando uma linguagem de programação específica. Dentre as plataformas mais conhecidas cita-se, Android, iOS e Windows Phone e suas respectivas linguagens de programação, JAVA, Objective-C e C++ ou C# [White 2103]. No desenvolvimento de aplicativos nativos é possível utilizar diversos recursos e funcionalidades do sistema operacional do dispositivo [Silva et al. 2015]. Uma das vantagens em se poder usar funcionalidades do SO[5] de um *smartphone*, é a possibilidade de enviar notificações para o aparelho do usuário, o que em aplicativos móveis *web* não pode ser feito.

Existem também os aplicativos híbridos, que utilizam tecnologias *web* para o desenvolvimento e ao mesmo tempo conseguem também utilizar recursos nativos dos dispositivos móveis [Andrade et al. 2013]. Os sistemas operacionais dos dispositivos móveis possuem em comum um tipo especial de *browser*, conhecido como *WebView*. Este navegador é acessível através de programação por código nativo, onde cada plataforma permite que seja aberta uma instância deste navegador, para que quando uma aplicação híbrida é aberta pelo usuário, esta seja executada dentro do *WebView*. Através dessa abordagem é possível criar aplicativos híbridos, utilizando HTML, CSS e JavaScript e ainda utilizar recursos nativos [Charland e Leroux 2011]. Para se obter uma aplicação multiplataforma, que consiga acessar recursos de *hardware* e *software* do dispositivo, o tipo de aplicação escolhido para o desenvolvimento do presente trabalho foi o de aplicação híbrida.

2.2. Trabalhos relacionados

Como trabalhos relacionados, foram encontradas três aplicações *mobiles* de serviços existentes no mercado para instituições de ensino. Uma breve descrição de ambas aplicações é abordada abaixo.

O primeiro trabalho relacionado foi o OutClass [Outclass 2016], que é um aplicativo *mobile* que conecta estudantes, pais, professores e instituições de ensino, afim de fornecer informações da vida acadêmica do aluno no *smartphone*. O Outclass não possui nenhuma funcionalidade ou serviço que envie informações sobre o acadêmico aos sistemas das instituições e escolas, ou seja, o aluno possui somente funcionalidades de

[4] URL: *Uniform Resource Locator*.
[5] SO: Sistema operacional.

consulta de dados. Instituições como COC Florianópolis e CEBRACORP utilizam o OutClass.

O VMobile [Verga Sistemas 2016] é um aplicativo voltado para a área da educação que pode ser integrado com o sistema de gestão educacional das instituições. A integração deste aplicativo é facilitada quando a instituição utiliza um sistema de gestão acadêmica específico, e com regras de negócio compatíveis com essa aplicação. Instituições como Unisinos e UNIFACEX utilizam o VMobile.

Já o EduApp [Eduapp 2016] é um aplicativo que não tem como foco principal a integração com sistemas de gestão educacional das instituições, no entanto, a integração pode ser realizada se houver esta possibilidade. A opção principal oferecida pela aplicação, onde é possível gerir o aplicativo sem a necessidade de integração, é por meio de um portal, que deve ser alimentado com dados e informações dos acadêmicos e consequentemente deve ser mantido e atualizado, tornando-se um sistema a parte.

3. Metodologia

Para o desenvolvimento do aplicativo, foi utilizado o Modelo Cascata como metodologia de desenvolvimento, apresentado por [Pressman 2011]. Segundo este autor, a etapa inicial que contempla o modelo em questão, é o levantamento de requisitos e necessidades por parte do cliente, avançando sequencialmente pelas fases de planejamento, modelagem, construção e implantação. Na etapa de implantação, de forma a verificar se os objetivos propostos foram alcançados, foi realizado um estudo de caso cujo objetivo foi integrar o aplicativo móvel desenvolvido a um sistema de gestão acadêmica real. Esse estudo de caso foi realizado na FACCAT, conforme as regras de negócio da mesma.

O desenvolvimento prático do aplicativo iniciou-se pela etapa de levantamento de requisitos e de necessidades. Esse levantamento teve início por meio de entrevistas com profissionais da área acadêmica da FACCAT. Essa análise teve como objetivo identificar quais funcionalidades seriam importantes existirem no aplicativo a ser desenvolvido, e que fosse comum a mais de uma instituição. Nessa etapa foi identificada a necessidade de existirem dois módulos, um dedicado ao público acadêmico e outro aos professores da instituição. Entretanto apenas o módulo para alunos foi contemplado no escopo do presente trabalho. Com base nisso, foi identificado que os atores do sistema são os alunos, tanto de graduação, pós-graduação e de cursos de extensão.

A partir desse estudo, pôde-se registrar os seguintes requisitos funcionais: (i) alteração de dados cadastrais; (ii) o aplicativo deve possibilitar a realização de inscrições em eventos e cursos de extensão; (iii) alteração e recuperação de senha; (iv) consultar dados financeiros; (v) consultar histórico escolar; (vi) consultar matrículas; (vii) consultar notas; (viii) consultar dados de horários de disciplinas; (ix) ativar/desativar o recebimento de notificações.

Como requisitos não funcionais, que [Pressman 2011] descreve como aqueles que não estão diretamente relacionados às funções específicas fornecidas pelo sistema, são listados: (i) ser operacionalizado em mais de uma plataforma móvel; (ii) garantir a integridade de dados e de processamento desses; (iii) o aplicativo deve ser integrável aos sistemas de gestão acadêmica das IES, permitindo que algumas informações sejam alteradas por meio do aplicativo.

Os casos de uso do *software* desenvolvido se diferem dos trabalhos relacionados no que tange a alteração de dados por meio do aplicativo. Nas outras aplicações apresentadas na Seção 2.2, e que são comuns a várias IES, a alteração de dados não ocorre, pois, cada IES possui regras de negócio específicas para validação dos dados informados. Para que fosse possível realizar a integração do aplicativo com mais de um sistema de gestão acadêmica, respeitando as regras de negócio de cada IES, uma API (*Application Programming Interface*) foi desenvolvida para a comunicação. A Figura 1 ilustra o diagrama de arquitetura da aplicação.

A API é utilizada como meio de comunicação entre o aplicativo móvel e o sistema de gestão acadêmica das IES e também é responsável pela troca de mensagens entre eles. Cada vez que o aplicativo é acessado, requisições HTTP (*request*) no formato *json* são do dispositivo móvel enviadas à API via *web service,* que por sua vez encaminha esses dados para o sistema da instituição. Por exemplo, na etapa de autenticação, o aluno informará suas credenciais por meio do aplicativo, que enviará esses dados para a API e, essa, encaminhará para o sistema da IES, onde de fato será feito a validação dos dados. Assim, nenhuma regra do sistema acadêmico da IES é comtemplada na API e sim no próprio sistema de gestão da IES. Logo, para que seja possível realizar a integração dos sistemas acadêmicos com o MobIES, é necessário que exista nos sistemas acadêmicos uma camada de API, a qual se comunicará com a API do MobIES. Dessa mesma forma são realizadas as validações de formulários e a exibição de mensagens aos usuários. Após o processamento dos dados no sistema acadêmico da IES, a API deste retorna uma mensagem à API do MobIES (de erro ou sucesso). Por sua vez, essa mensagem é direcionada para a aplicação móvel.

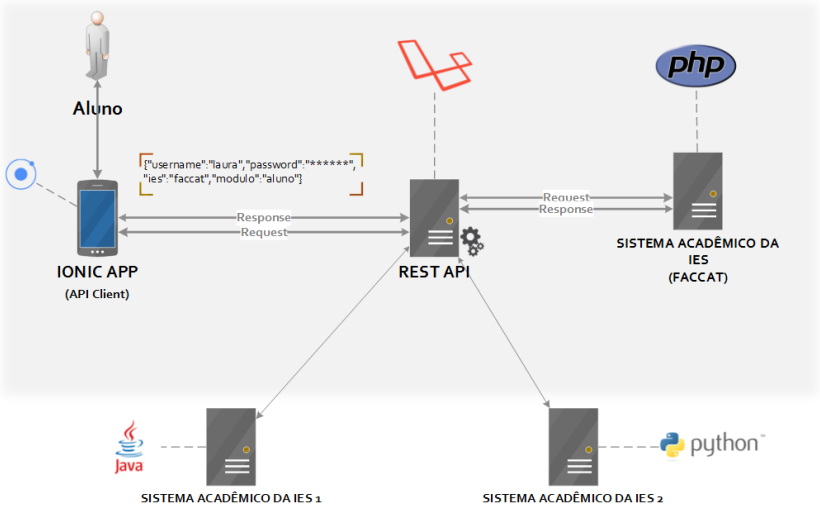


Figura 1. Diagrama de arquitetura da aplicação

A fase de desenvolvimento se dividiu em três etapas. Na primeira houve o desenvolvimento do aplicativo. Na segunda etapa, foi construída uma aplicação de *backend*[6]. Por fim, o sistema acadêmico da FACCAT foi alterado para permitir a sua integração com o *backend* do aplicativo móvel.

[6] *Backend*: código rodado do lado do servidor, também conhecido como *server-side*.

O desenvolvimento do aplicativo móvel foi realizado com a tecnologia IONIC. O IONIC é um *framework front-end*[7] de código fonte aberto utilizado para o desenvolvimento de aplicativos híbridos, baseado em tecnologias *web* HTML, CSS e JavaScript (AngularJS). Este *framework* utiliza as plataformas Cordova/Phonegap para distribuir o aplicativo para os dispositivos móveis [Ionic 2016]. Para o desenvolvimento da API, o *framework* Laravel foi utilizado. A API desenvolvida é do tipo REST (*REpresentational State Transfer*), descrita em [Abeysinghe 2008]. A troca de mensagens entre API, aplicativo móvel e o sistema acadêmico da IES é feito através de *web services* Gomes [2014], onde o formato de mensagem escolhido para ser utilizado foi o *json*. A camada de transporte utilizada para a troca de mensagens é o HTTP e os métodos utilizados foram o GET e POST. Para realizar a autenticação dos usuários com os sistemas envolvidos, a biblioteca OAuth 2 foi utilizada.

4. Resultados

Como resultados, obteve-se uma aplicação híbrida batizada de MobIES, onde o principal benefício por parte da IES, é o de oferecer aos alunos serviços institucionais através de um meio de comunicação atualmente importante e presente no cotidiano da grande maioria dos estudantes. Com o MobIES a instituição consegue atender seu público de uma forma alternativa, além do atendimento convencional. Já por parte do acadêmico, além da vantagem de mobilidade de acesso aos dados, também há um ganho em otimização do processo de acompanhamento de informações. Alguns dos serviços institucionais e funcionalidades são descritos abaixo.

Ao acessar o MobIES, inicialmente na tela de login, o aluno deverá informar qual a sua IES, e a partir desta informação, a aplicação consegue realizar todas as demais funcionalidades existentes, inclusive identificar o aluno. Após a realização do login, o aluno será direcionado para uma tela de "Matrícula" que exibe as matrículas ativas do aluno e suas respectivas disciplinas do semestre atual. A partir dessa tela, o aluno tem a opção de ver informações detalhadas das disciplinas, como nome do professor e sala de aula, assim como informações sobre suas notas e provas.

O aplicativo possui um menu lateral (Figura 2) que exibe ao acadêmico todas as funcionalidades existentes, inclusive as descritas acima. A funcionalidade de "Histórico escolar" exibe dados das disciplinas já cursadas pelo aluno, médias finais e situação, os "Horários" exibe os dias, salas e horárias das disciplinas dos respectivos cursos do aluno no semestre atual e a funcionalidade "Financeiro" exibe ao aluno, caso ele possua, os débitos em aberto, o valor dos débitos, data de vencimento, descontos, etc.

Existem funcionalidades que permitem ao acadêmico realizar alterações, por exemplo, a de "Dados cadastrais". Nessa, o aluno pode alterar e atualizar os dados de seu cadastro junto à instituição. Outro exemplo é a funcionalidade de "Inscrições em eventos", onde é possível visualizar todos os eventos/cursos que estão ocorrendo na instituição e realizar inscrições nesses eventos. Após efetivada a inscrição, uma tela de consulta de inscrições é disponibilizada e depois da ocorrência do evento, caso o aluno possua a frequência necessária, é possível solicitar através do aplicativo o envio do certificado por e-mail.

[7] *Front-end*: o que é exibido para o usuário, responsável por coletar dados de entrada desse.

Estas funcionalidades de alterações e as demais existentes, como “Troca de senha” e “Recuperação de senha”, alteram os dados no sistema acadêmico da instituição. No momento em que o aluno realiza a alteração no aplicativo, requisições HTTP (*request*) no formato *json* são enviadas à API via *web service* que envia os dados para o sistema da instituição. Todas as validações de dados e de formulários, inclusive nas telas exibidas ao aluno são definidas pelo sistema da instituição. Após as validações, uma resposta (*response*) é entregue ao aplicativo, informando se a alteração foi concluída ou não.

Figura 2. Funcionalidades do app (menu)

5. Conclusão

O presente artigo apresentou o desenvolvimento de um aplicativo *mobile*, denominado MobIES, o qual teve como finalidade ser um aplicativo comum a mais de uma instituição, além disso o MobIES tem o objetivo de oferecer aos alunos uma forma alternativa de atendimento e fornecimento de dados e informações, respeitando as regras de negócio de cada IES.

Durante o desenvolvimento deste trabalho, foi identificado que existem poucos aplicativos no mercado que fornecem esse tipo de serviço e que as possíveis causas para isso, são as regras de negócio estabelecidas pelas instituições. Identificou-se também, que a maioria das IES ainda não possuem uma aplicação móvel disponível ao público acadêmico e que o fornecimento desse tipo de serviço atualmente é importante para as instituições e significativo para os alunos.

O MobIES foi desenvolvido utilizando tecnologia híbrida, que permite a ele funcionar em diferentes sistemas operacionais móveis, mais especificamente em *Android* e *iOS*. Afim de facilitar a integração do aplicativo com mais de uma instituição, foi desenvolvido uma API, responsável pela comunicação do sistema de gestão da IES com o aplicativo móvel, ou seja, não existe nenhum tipo de comunicação direta entre aplicação *mobile* e o sistema acadêmico da instituição. No momento, a aplicação encontra-se em fase de testes, e pretende-se até fevereiro de 2017 estar à disposição para o público acadêmico da FACCAT. Esses testes, ainda não foram realizados com o público acadêmico, apenas pelos desenvolvedores da área de TI dessa instituição.

A solução desenvolvida conseguiu atender de forma satisfatória os objetivos descritos neste artigo. Ainda é necessário concluir o envio de notificações e possibilitar que o próprio aluno configure se quer ou não receber este tipo de aviso. A funcionalidade em questão, irá possibilitar aos acadêmicos um melhor acompanhamento quando suas notas são postadas no sistema acadêmico da IES. Pretende-se também como trabalhos futuros, desenvolver o módulo de docentes e identificar quais as funcionalidades que seriam benéficas a este público.

Referências

Andrade, A. W., Agra, R. e Malheiros, V. (2013) "Estudos de caso de aplicativos móveis no geverno brasileiro", Serviço Federal de Processamento de Dados, SERPRO, Brasília.

Charland A. e Leroux B. (2011) "Mobile Application Development: Web vs. Native", Communications of the ACM, p. 49-53.

Eduapp. (2016), Disponível em: http://eduapp.com.br, acesso em agosto de 2016.

Gomes, D. A. (2014) "Web Services SOAP em JAVA: Guia prático para o desenvolvimento de web services em Java", Novatec, 2ª edição.

Gomes, M., Fávero, N. C. e Lucas, C. (2014) "Gerenciamento do Relacionamento com o Estudante no ensino superior", Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica, Centro Universitário de Franca, Uni-FACEF, p. 76-98, v. 4, n. 1.

Ionic. (2016) "Ionic Documentation Overview", Disponível em: http://ionicframework.com/docs/overview, acesso em setembro de 2016.

Martins, C. de S., Antônio, A. L. T. de e Oliveira, C. A. de. (2013) "Os desafios para a mobilização de aplicações baseadas em plataforma Web", III Escola Regional de Informática, Regional Norte, Boa Vista.

Nascimento, H. J. (2013) "Um projeto de aplicativo móvel para entender o conceito de função matemática", XVII Encontro Nacional de Estudantes de Pós-Graduação em Educação Matemática, Instituto Federal do Espírito Santo.

Nonnenmacher, R. F. (2012) "Estudo do comportamento do consumidor de aplicativos móveis", Departamento de Ciências Administrativas, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Outclass (2016), Disponível em: http://outclassapp.com, acesso em Agosto de 2016.

Presman, R. S. (2011) "Engenharia de Software: Uma Abordagem Profissional", Bookman Mc Graw Hill, 7ª edição, Porto Alegre.

Silva, L. L. B. da, Pires, D. F. e Neto, S. C. (2015) "Desenvolvimento de Aplicações para Dispositivos Móveis: Tipos e Exemplos de Aplicação na plataforma iOS", II Workshop de Iniciação Científica de Franca, Goiânia.

Verga Sistemas. (2016), Disponível em: http://verga.com.br/vmobile, acesso em agosto de 2016.

White, J. (2013) "Going native (or not): Five questions to ask mobile application developers", Australian Medical Journal, http://dx.doi.org/10.4066/AMJ.2013.1576.

Organização **Realização** **Apoio**

ISBN: 978-85-7669-356-7 (online)