Realização:

SBC
Sociedade Brasileira de Computação

SBSI 2017

XIII Simpósio Brasileiro de Sistema de Informação - Lavras/MG

SISTEMAS DE INFORMAÇÃO PARA GOVERNANÇA COM ENVOLVIMENTO E PROTAGONISMO DO CIDADÃO

X WORKSHOP DE TESES E DISSERTAÇÕES EM SISTEMAS DE INFORMAÇÃO

UFLA - Lavras, MG - 5 a 8 de Junho de 2017

Organização: UFLA Universidade Federal de Lavras

Afiliado à:

Cooperação: acm

Fomento: CAPES 65 anos

Patrocínio Prata: SISTEMA INFORMÁTICA

Apoio: compJÚNIOR; Centro Acadêmico de Sistemas de Informação; INNERVISION; 30 anos SQUADRA Tecnologia

X Workshop de Teses e Dissertações em Sistemas de Informação (WTDSI)

Evento integrante do XIII Simpósio Brasileiro de Sistemas de Informação

De 5 a 8 de Junho de 2017

Lavras – MG

ANAIS

Sociedade Brasileira de Computação – SBC

Organizadores

Paulo Afonso Parreira Júnior

Rita Suzana Pitangueira Maciel

Heitor Augustus Xavier Costa

Juliana Galvani Greghi

Realização

DCC/UFLA - Departamento de Ciência da Computação/Universidade Federal de Lavras

Sociedade Brasileira de Computação – SBC

Patrocínio Institucional

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

Ficha Catalográfica preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Workshop de Teses e Dissertações em Sistemas de Informação (13. : 2017 : Lavras, MG)

Anais do X Workshop de Teses e Dissertações em Sistemas de Informação WTDSI 2017; organizadores: Paulo Afonso Parreira Júnior, Rita Suzana Pitangueira Maciel, Heitor Augustus Xavier Costa, Juliana Galvani Greghi; realização: Universidade Federal de Lavras - UFLA, Sociedade Brasileira de Computação – SBC – Lavras: UFLA, 2017.

91 p. : il.
Bibliografias.
Disponível em: http://sbsi2017.dcc.ufla.br/anais.html.
ISBN: 978-85-7669-377-2

1. Sistemas de recuperação da informação Congressos. 2. Tecnologia serviços de informação Congressos. 3. Internet na administração pública Congressos. I. Parreira Júnior, P. A. II. Maciel, R. S. P. III. Costa, H. A. X. IV. Greghi, J. G. V. Universidade Federal de Lavras. VI. Sociedade Brasileira de Computação. VII. Título.

CDD-658.4038
-658.4038011

X WTDSI

X Workshop de Teses e Dissertações em Sistemas de Informação (WTDSI)
Evento integrante do XIII Simpósio Brasileiro de Sistemas de Informação (SBSI)
5 a 8 de Junho de 2017
Lavras, Minas Gerais, Brasil.

Comitês

Coordenação Geral do SBSI 2017
Heitor Augustus Xavier Costa (UFLA)
Juliana Galvani Greghi (UFLA)

Coordenação do Comitê de Programa do WTDSI 2017
Paulo Afonso Parreira Júnior (UFLA)
Rita Suzana Pitangueira Maciel (UFBA)

Comissão Especial de Sistemas de Informação
Clodis Boscarioli (UNIOESTE)
Renata Araujo (UNIRIO)
Andrea Magalhães Magdaleno (UFF)
Claudia Cappelli (UNIRIO)
Patricia Vilain (UFSC)
Raul Sidnei Wazlawick (UFSC)
Sean Wolfgand Matsui Siqueira (UNIRIO)
Valdemar Vicente Graciano Neto (UFG)

Comitê de Programa Científico do WTDSI 2017
Adolfo Duran (UFBA)
Ana Carolina Inocêncio (UFG/Jataí)
Ana Patricia Fontes Magalhaes Mascarenhas (UNEB)
Ana Paula Vilela (UFG/Jataí)
Antonio Resende (UFLA)
Bruno Silva (UFLA)
Carlos Santos Jr. (UnB)
Celia Ralha (UnB)
Claudia Cappelli (UNIRIO)
Daniel de Oliveira (UFF)
Davi Viana (UFMA)
Edmir Prado (USP)
Fatima Nunes (EACH-USP)
Fernanda Lima (UnB)
Fernanda Campos (UFJF)
Fernanda Baião (UNIRIO)
Flavia Santoro (UNIRIO)
Jairo Souza (UFJF)
João Porto de Albuquerque (University of Warwick)
José David (UFJF)
Leonardo Azevedo (IBM; UNIRIO)
Lucineia Heloisa Thom (UFRGS)
Marcelo Morandini (USP)
Marcos Wagner Souza Ribeiro (UFG/Jataí)
Maria Luiza (UFRJ)
Melise Paula (UNIFEI)
Rafael Durelli (UFLA)
Rodrigo Santos (UNIRIO)
Sean Siqueira (UNIRIO)
Tatiane Nogueira (UFBA)
Vera Werneck (UNIRIO)

Realização
DCC/UFLA - Departamento de Ciência da Computação/Universidade Federal de Lavras

Promoção
Sociedade Brasileira de Computação – SBC

Patrocínio Institucional
CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Apoio

Patrocínio Prata
Sistema Informática

Apoio
Centro Acadêmico de Sistemas de Informação – UFLA
CompJúnior - UFLA
INNERVISION
SQUADRA Tecnologia

Apresentação

O Simpósio Brasileiro de Sistemas de Informação (SBSI) é um evento para a apresentação de trabalhos científicos e discussão de temas relevantes na área de Sistemas de Informação, aproximando estudantes, pesquisadores, profissionais e empresários da comunidade de Sistemas de Informação. As pesquisas em Sistemas de Informação combinam aspectos multidisciplinares das áreas da Ciência da Computação, Matemática, Ciência da Informação, Administração, Comportamento Organizacional, entre outras. A aplicação dos diversos domínios do conhecimento na busca por soluções aos problemas envolvendo os Sistemas de Informação é o que motiva a realização deste evento.

O Workshop de Teses e Dissertações em Sistemas de Informação (WTDSI) é um fórum dedicado à apresentação e discussão de trabalhos de mestrado e de doutorado em Sistemas de Informação, desenvolvidos nos programas de pós-graduação no Brasil. O seu objetivo é propiciar um ambiente construtivo para discussões, em que os alunos possam ter uma avaliação dos seus trabalhos por pesquisadores experientes em Sistemas de Informação.

Em sua 10ª. edição, o WTDSI recebeu um total de 35 submissões, sendo 4 delas sobre propostas de doutorado e 31 de mestrado. Das 35 submissões, 60% advinham do estado de São Paulo, em particular, da instituição EACH-USP; 25,7% vieram de Universidades do Nordeste do país, mais especificamente, da Bahia (UFBA) e de Pernambuco (UFPE); os 14,3% restantes vieram do Rio de Janeiro e do Rio Grande do Sul.

Cada trabalho submetido foi avaliado por, no mínimo, três membros do comitê de programa e, ao final do processo de revisão, 22 artigos foram aceitos para apresentação no WTDSI. Os trabalhos selecionados apresentam um panorama da pesquisa nos programas de pós-graduação em SI no Brasil, envolvendo temas e problemáticas atuais e relevantes, que foram divididas nas seguintes sessões de apresentação: Sistemas Inteligentes e de Apoio à Decisão em SI, Governo, Educação e Integração de Aplicações Corporativas, Interação Humano-Computador e Aspectos Humanos e Sociais em SI e Gestão de Processos, Metodologias e Abordagens para SI.

A coordenação agradece aos autores dos trabalhos e seus orientadores, por prestigiarem o WTDSI 2017; aos membros do comitê de programa, pelo tempo dedicado às revisões; e à organização geral do SBSI, por todo o suporte oferecido. Que o WTDSI possa contribuir positivamente para o bom andamento dos trabalhos a ele submetidos...

Lavras, Junho de 2017.

Paulo Afonso Parreira Júnior (UFLA) e Rita Suzana Pitangueira Maciel (UFBA)
Coordenação do WTDSI 2017

Biografia dos Coordenadores do Comitê de Programa do WTDSI 2017

Possui graduação em Ciência da Computação pela Universidade Federal de Lavras (2009), mestrado (2011) e doutorado (2015) em Ciência da Computação (área: Engenharia de Software) pelo Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos (PPGCC/UFSCar). Atualmente é professor do Departamento de Ciência da Computação da Universidade Federal de Lavras (DCC/UFLA) e integrante do Grupo de Pesquisa em Engenharia de Software (PEQS). Tem experiência na área de Ciência da Computação, com ênfase em Engenharia de Software, atuando principalmente em: Manutenção de Software, Engenharia de Requisitos, Desenvolvimento de Software Orientado a Objetos, Desenvolvimento de Software Orientado a Aspectos e Informática na Educação.

Rita Suzana é atualmente é professora Associada do Departamento de Computação da Universidade Federal da Bahia, onde atua na graduação e pós-graduação. Possui mestrado em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (1995) e doutorado em Ciência da Computação pela Universidade Federal de Pernambuco (2005), tendo realizado Pós Doutorado na University of Waterloo-Canadá (2014). Sua área de pesquisa possui forte concentração na Engenharia de Software e como esta pode apoiar mais especificamente o desenvolvimento de Sistemas de Informação e Sistemas Colaborativos.

Sumário

Um Algoritmo de Aprendizado por Reforço para Ranking de Relacionamentos na Web Semântica

Alternative Title: A Reinforcement Learning Algorithm to Relationship Ranking in the Semantic Web

Paulo M. F. dos Santos
Escola de Artes, Ciências e Humanidades - USP
paulofranco@usp.br

Karina V. Delgado (Orientadora)
Escola de Artes, Ciências e Humanidades - USP
kvd@usp.br

RESUMO

O ato de realizar pesquisas na *Web* tem sido o mesmo por anos, nessas pesquisas o usuário realiza uma consulta composta de termos relacionados ao que ele deseja encontrar, e o motor de busca é responsável por encontrar as melhores respostas àquela consulta. Frequentemente, existe muito mais na cabeça do usuário ao fazer sua consulta que não é transmitido, mas que ele espera que o motor de busca seja capaz de inferir. Isso leva a resultados que são relacionados à sua consulta, mas não aos seus interesses. Uma maneira de resolver esse problema foi introduzido através da busca semântica, que visa a permitir que os dados disponíveis na *Web* tenham um sentido, ou seja, uma semântica. Diversas abordagens de busca na *Web* Semântica têm sido propostas e implementadas nos últimos anos, bem como abordagens para classificação (*ranking*) de resultados. Este projeto de mestrado tem por objetivo aplicar aprendizado por reforço em um motor de busca da *Web* Semântica como uma técnica para ranking de relacionamentos, de forma a se obter um algoritmo personalizado e com capacidade de evoluir com o uso contínuo.

Palavras-Chave

Web Semântica, Linked Data, Ranking de Relacionamentos, Aprendizado por Reforço.

ABSTRACT

The act of searching the Web has been the same for years. The user inputs a query and the search engine is responsible for finding the best matches to that query. Often, there is subjective information that the user can not transmit when making his query, but he expects that the search engine will infer. This leads to results that are query-related, but not user-interest related. One way of mitigating this problem was the introduction of the Semantic Web, which aims to allow that the data available on the Web have a meaning. Many search approaches on Semantic Web search have been proposed and implemented, as well as solutions to rank the results. This masters project aims to apply reinforcement learning on a search engine for the Semantic Web as a technique to rank the relationships, to obtain a persinalized algorithm with the capacity to evolve with continuous use.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•Information systems → Learning to rank; •Theory of computation → Reinforcement learning;

Keywords

Semantic *Web*, Linked Data, Relationship Ranking, Reinforcement Learning.

1. INTRODUÇÃO

Com a popularização da internet nos anos 90, a quantidade de páginas disponíveis na *World Wide Web* (abreviada informalmente para *Web*) cresceu em um nível exponencial. Cada página da Web representa uma fonte de conhecimento e interesse, e essas páginas e seus conteúdos vêm sendo criados e armazenados da mesma forma ao longo dos anos. Motores de busca, que realizam pesquisa sobre a Web, vêm se tornando cada vez mais poderosos, com seus algoritmos otimizados para recuperar a maior quantidade de informação, na menor quantidade de tempo possível. No entanto, devido à dificuldade ou impossibilidade de se introduzir informações subjetivas nas consultas, pode ser que uma grande quantidade dos resultados encontrados pelos motores de busca tenham baixa relevância para o usuário [13]. Esse fenômeno é observado em motores de busca baseados exclusivamente em consultas por palavras-chave.

Para mitigar esse problema, foi criada uma proposta de evolução da Web, denominada *Web Semântica*. A Web Semântica é uma extensão da Web que introduz padrões para o compartilhamento de dados na rede, de forma a permitir que os dados disponíveis na Web passem a ter um sentido, isto é, uma semântica. Os dados podem assim ser distinguidos pelas máquinas, que agora trabalham com o próprio conhecimento acerca das informações, ao invés de trabalhar apenas com textos, facilitando a recuperação automática de informações.

A área de pesquisa em Web Semântica tem evoluído significativamente, com o objetivo de mudar a maneira em que as informações são organizadas, armazenadas e recuperadas

na Web. Motores de busca semântica têm sido desenvolvidos na última década, com o foco de recuperar informações que sejam mais relevantes para o usuário. Mais recentemente, os dados na Web Semântica mudaram da maneira caótica em que eram organizados e distribuídos para uma maneira mais fácil e concisa, denominada *Linked Data*. Linked Data são uma série de boas práticas desenvolvidas para organizar os dados na Web em entidades e relacionamentos, interconectando-os em estruturas de grafos [4]. Atualmente, a organização e disponibi-lização desses dados se dá através do *Resource Description Framework* (RDF), um modelo de organização de dados que se aproxima de conceitos de modelagem clássicos de bancos de dados.

Um dos fatores importantes na recuperação de informação é como organizar e apresentar os resultados das buscas. Para tal, é necessário escolher uma classificação subjetiva de importância para esses dados. Essa organização dos resultados através de um critério subjetivo é denominada *ranking*. No caso da Web Semântica, existem diversos critérios que podem ser adotados para formar diferentes rankings das informações que a constituem. É possível organizar e classificar seus dados através da importância relativa das propriedades RDF, ou através de critérios como popularidade das entidades, raridade dos conceitos, entre outros.

Jindal et al. [11] classificam as abordagens de ranking para a Web Semântica em três categorias: ranking de entidades, ranking de relacionamentos e ranking de documentos.

Os algoritmos de ranking de entidade são aqueles que têm como objetivo recuperar resultados baseados nas entidades de interesse fornecidas como entrada pelos usuários ao motor de busca. Esses algoritmos encontram a proximidade da entidade com as entidades vizinhas baseados na quantidade de relacionamentos entre elas [11]. Alguns exemplos de estudos que fazem esse tipo de ranking são [20, 14, 10].

Os algoritmos de ranking de relacionamento são aqueles que focam na importância relativa dos relacionamentos entre as entidades, com relação ao contexto da consulta do usuário. Neste projeto de mestrado, estamos interessados nesta abordagem de ranking.

Os algoritmos de ranking de documentos são os que buscam documentos com o mais relevante e completo conjunto de entidades de interesse, bem como o conjunto mais relevante de relacionamentos entre essas entidades [11]. Entre os trabalhos que seguem essa abordagem estão [1, 13, 6].

2. APRESENTAÇÃO DO PROBLEMA

Os relacionamentos são a parte mais importante da semântica pois é através dele que a informação ganha significado, torna-a mais compreensível e até mesmo fornece novos conhecimentos, às vezes até inesperados, acerca dela [1]. Na literatura, encontra-se diversas propostas de algoritmos de ranking para relacionamentos semânticos, desde algoritmos baseados em métricas semânticas [2, 12, 17] até probabilísticos [5]. Porém existem poucas propostas de algoritmos que apresentem capacidade de evolução ao longo do tempo.

Em uma revisão sistemática recente dos últimos dez anos elaborada pelos autores, encontrou-se apenas um algoritmo com essa capacidade, o *RankBox*[7]. Para tal, os autores aplicaram técnicas de aprendizado de máquina como forma de automatizar esse processo. Foram exploradas duas técnicas distintas em [7] e [8]. Uma utilizando *support vector machine* (SVM) e a outra utilizando *linear discriminant analysis* (LDA). A técnica baseada em SVM precisava de ajustes manuais, e não era capaz de mudar com o tempo. Já a técnica baseada em LDA veio como evolução da proposta anterior, criando um modelo sem necessidade de definição de parâmetros por parte dos usuários, bem como permitiu a evolução contínua do algoritmo. Apesar disso, não encontrou-se nos últimos anos outra técnica que apresentasse essas características.

Acredita-se que ser capaz de evoluir com o tempo, adaptando seus parâmetros ao gosto do usuário, que por essência é um ser cujos desejos são também mutáveis ao longo do tempo, é de suma importância para que o usuário fique mais satisfeito com a ordenação dos resultados devolvidos. Desta forma, outras técnicas de ranking com essa capacidade devem ser exploradas.

3. PROPOSTA DE SOLUÇÃO

A abordagem proposta busca utilizar as vantagens da automatização do processo de personalização do algoritmo através do uso de aprendizado de máquina, e explorar uma solução diferente das encontradas na literatura sobre Web Semântica, neste caso, com o uso de aprendizado por reforço para realizar o ranking dos relacionamentos semânticos.

Algoritmos de aprendizado de máquina já vem sendo aplicados na Web tradicional para criar rankings personalizados das buscas dos usuários. Esses algoritmos são chamados de Learn-to-Rank (LTR). Entre eles, técnicas de aprendizado por reforço, usadas por exemplo em [9], [15] e [19] podem ser adaptadas para a Web Semântica, com o intuito de criar um algoritmo de ranking personalizado e evolutivo.

Aprendizado por reforço é uma técnica de aprendizado de máquina não supervisionada, ou seja, que não necessita de conhecimento prévio sobre o ambiente que se deseja explorar. Em particular, um caso especial do aprendizado por reforço, os chamados problemas *bandit*, podem ser adaptados para o problema do ranking de relacionamentos na Web Semântica. Assim, pretende-se utilizar um *multi-armed bandit* como proposta de solução.

A Figura 1 representa o funcionamento do sistema proposto. Um usuário realiza uma busca por palavras-chave através da interface do sistema, e o sistema utiliza a linguagem SPARQLeR para realizar a consulta dos dados semânticos. O sistema cria então uma lista de associações relacionadas à busca do usuário, e envia a lista para a parte do algoritmo responsável por fazer o ranking. O ranking utilizará uma quantidade de *arms* para o problema bandit, em que cada *arm* ficará responsável por escolher um resultado para ser inserido em uma posição da lista ordenada final (por exemplo, o *Arm* 1 escolhe o primeiro elemento da lista, o *Arm* 2 o segundo, etc). Após criada a lista ordenada, esta é devolvida para o usuário através da interface. O usuário então fornecerá um *feedback* , seja de maneira implícita (clicando em um link de associação), ou explícita (gostando ou não gostando de uma associação em uma determinada posição). Esse *feedback* é enviado para o sistema, que recompensará ou punirá cada *arm* e ajustará os parâmetros para a próxima consulta. Esse ciclo de busca-ranking-feedback contínuo é responsável pelo treinamento evolutivo do algoritmo.

Algumas políticas podem ser adotadas no algoritmo, como busca por recompensas imediatas, planejamento de recompensas a longo prazo, diversificação dos resultados, entre outras, cada uma com suas vantagens e desvantagens. Note que, com o tempo, o usuário precisará dar menos estímulos

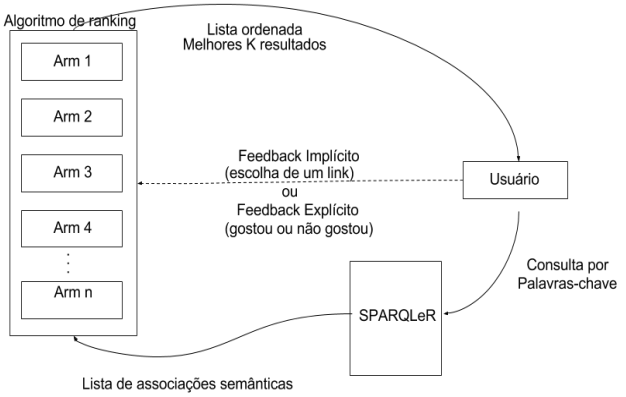


Figura 1: Funcionamento do sistema proposto.

explícitos ao sistema, pois este deverá ter se tornado capaz de entender as suas preferências.A solução difere de algoritmos clássicos de LTR pelo fato de que, em sua grande maioria, eles precisam de um especialista que classifique os dados de treinamento previamente ao uso. Na abordagem proposta, o treinamento dos dados é feito de forma contínua ao longo do uso.

4. AVALIAÇÃO DA SOLUÇÃO

Algoritmos para resolver o problema *bandit*, como o *Thompson Sampling* [16], o *Upper Confidence Bound*[3] e o *Adbandit*[18] serão implementados, para se analisar qual apresenta melhor desempenho no ranking de relacionamentos. Em seguida, serão feitas comparações com o algoritmo *RankBox* [7], que é o algoritmo mais recente na literatura que apresenta a capacidade de evolução com o tempo. Pretende-se utilizar as mesmas bases de dados, ou seja, a Freebase linked-open-data[1], com suas 340 mil instâncias e 500 mil relações semânticas.

O sistema será avaliado por 10 usuários e as métricas de avaliação a serem usadas são acurácia, tempo computacional e quantidade de iterações até atingir um nível de precisão aceitável pelo usuário.

5. ATIVIDADES REALIZADAS

Foi realizada uma revisão sistemática da literatura sobre ranking de relacionamentos da Web Semântica. Nesta revisão, encontrou-se apenas um algoritmo com a capacidade de evolução da função de ranking ao longo do tempo. Foi escolhido este algoritmo para ser implementado e comparado com a solução proposta.

As próximas etapas envolverão a implementação do algoritmo encontrado na literatura, o projeto e implementação do algoritmo proposto, e as análises dos resultados obtidos.

6. CONCLUSÃO

A Web Semântica, por se tratar de uma proposta de evolução da Web tradicional, necessita de estudos que viabilizem sua adoção e despertem o interesse dos usuários. Algoritmos que facilitem a navegação e o uso desse novo modo de organizar os dados são essenciais para esse objetivo. Este projeto de mestrado busca propor uma técnica que permita aos usuários realizarem buscas na Web Semântica em que o interesse nos relacionamentos semânticos seja o elemento mais importante através do uso do aprendizado por reforço. Desta forma, o usuário pode ter um sistema de busca personalizado para seus gostos, com a possibilidade de evolução através do uso contínuo.

[1]http://schemaviz.freebaseapps.com/?domain=/fictional_universe

7. REFERÊNCIAS

[1] B. Aleman-Meza, I. B. Arpinar, M. V. Nural, and A. P. Sheth. Ranking documents semantically using ontological relationships. In *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing*, pages 299–304. IEEE, 2010.

[2] B. Aleman-Meza, C. Halaschek-Wiener, A. Sheth, I. B. Arpinar, and C. Ramakrishnan. Ranking complex relationships on the semantic web. *IEEE Internet Computing*, pages 37–44, 2005.

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, Jan. 2003.

[4] T. Bernes-Lee. Linked data - design issues. http://www.w3.org/DesignIssues/LinkedData.html, 2006. Acessado em 31-01-2017.

[5] S. Bhatia, A. Goel, E. Bowen, and A. Jain. Separating wheat from the chaff – a relationship ranking algorithm. *The Semantic Web: ESWC 2016 Satellite Events*, pages 79–83, 2016.

[6] P. Castells, M. Fernandez, and D. Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE transactions on knowledge and data engineering*, 19(2):261–272, 2007.

[7] N. Chen and V. Prasanna. Rankbox: An adaptive ranking system for mining complex semantic relationships using user feedback. In *Proceedings of the IEEE 13th International Conference on Information Reuse and Integration*, pages 77–84, 2012.

[8] N. Chen and V. K. Prasanna. Learning to rank complex semantic relationships. *International Journal on Semantic Web and Information Systems*, 8(4):1–19, Oct. 2012.

[9] A. Grotov and M. de Rijke. Online learning to rank for information retrieval: Sigir 2016 tutorial. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1215–1218. ACM, 2016.

[10] A. Hogan, S. Decker, and A. Harth. Reconrank: A scalable ranking method for semantic web data with context. In *Proceedings of second international workshop on scalable semantic web knowledge base systems*, 2006.

[11] V. Jindal, S. Bawa, and S. Batra. A review of ranking approaches for semantic search on web. *Information Processing and Management*, 50(146):416–425, 2014.

[12] S. A. Kareem and P. M. Barnaghi. A context-aware ranking method for the complex relationships on the semantic web. In *International Conference on Advanced Language Processing and Web Information Technology*, pages 129–134. IEEE, 2007.

[13] F. Lamberti, A. Sanna, and C. Demartini. A relation-based page rank algorithm for semantic web

search engines. *IEEE Transactions on Knowledge and Data Engineering*, 21(1):123–136, 2009.

[14] X. Ning, H. Jin, and H. Wu. Rss: A framework enabling ranked search on the semantic web. *Information Processing & Management*, 44(2):893 – 909, 2008.

[15] J. Rennie and A. McCallum. Using reinforcement learning to spider the web efficiently. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 335–343. Morgan Kaufmann Publishers Inc., 1999.

[16] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[17] A. K. Thushar. An RDF approach for discovering the relevant semantic associations in a social network. In *16th International Conference on Advanced Computing and Communications*, pages 214–220, 2008.

[18] F. S. Truzzi, V. F. da Silva, A. H. R. Costa, and F. G. Cozman. Adbandit: a new algorithm for the multi-armed bandits. In *Encontro Nacional de Inteligência Artificial(ENIA)*, 2013.

[19] V.Derhami, J.Paksima, and H. Khajeh. Web pages ranking algorithm based on reinforcement learning and user feedback. *Journal of AI and Data Mining*, 3(2):157–168, 2015.

[20] W. Wei, P. Barnaghi, and A. Bargiela. Rational research model for ranking semantic entities. *Information Sciences*, 181(13):2823 – 2840, 2011.

Algoritmo de Aprendizado por Reforço para Resolução de Processos de Decisão Markovianos Sensíveis ao Risco

Alternative Title: Reinforcement Learning Algorithm to Solve Risk Sensitive Markov Decision Processes

Igor Oliveira Borges
Escola de Artes, Ciências e Humanidades - USP
igor.borges@usp.br

Karina V. Delgado (Orientadora)
Escola de Artes, Ciências e Humanidades - USP
kvd@usp.br

RESUMO

Problemas de decisão podem ser modelados usando Processos Markovianos de Decisão (MDP). Existe uma extensão de MDP que permite lidar com risco, o MDP Sensível ao Risco (RSMDP). Porém, existem poucos algoritmos de aprendizado por reforço seguro para RSMDP, e seu desempenho precisa ser melhorado para garantir a aplicabilidade a problemas mais complexos. A resolução de RSMDP adotada é realizada por meio critério de otimização com enfoque na abordagem de utilidade exponencial, o artigo mostra alguns desafios e possíveis soluções desta modelagem para auxiliar no desenvolvimento de algoritmos de Aprendizado por Reforço Sensível ao Risco. O objetivo desta pesquisa é desenvolver uma estratégia para tornar algoritmos de aprendizado por reforço mais eficientes para RSMDPs. Pretende-se que essa estratégia de aprendizado por reforço seja superior as estratégias disponíveis na literatura. O algoritmo será testado em dois domínios, o primeiro é um tipo do *GridWorld*, e o outro é um novo domínio de teste simulado inspirado no jogo de futebol discretizado em duas dimensões. Será comparada a curva de convergência do aprendizado e a qualidade da política obtida no algoritmo proposto com outras implementações disponíveis na literatura, considerando os mesmos parâmetros e processo de exploração. Por fim será discutido o impacto que atitudes ao risco distintas a neutralidade podem trazer na resolução de processos de decisão markovianos.

Palavras-Chave

Processo de Decisão de Markov, Processo de Decisão de Markov Sensível ao Risco, Utilidade Exponencial

ABSTRACT

Decision problems can be modeled using Markovian Decision Processes (MDP). There is an MDP extension that can deal with risk, called Risk Sensitive MDP (RSMDP). However, there are so few safe reinforcement learning algorithms for RSMDP, and their performance needs to be improved to ensure applicability in problems more complex. The adopted RSMDP resolution is performed through an optimization criterion focused on exponential utility approach, this article shows some challenges and possible solutions for defined model to help in development of Risk Sensitive Reinforcement Learning algorithms. The goal of this research is to develop a strategy to make more efficient reinforcement learning algorithms for RSMDPs. This reinforcement learning strategy intends to be superior compared to other strategies available in literature. The algorithm will be tested into two domains, first is a type of *GridWorld* domain, and other is a new test simulated domain inspired in a discrete two-dimensional soccer game. Convergence curve of learning and quality of obtained policy in proposed algorithm will be compared to other available implementations in literature, using same parameters and exploration process. Finally, this article will have a discussion about the impact that risk attitudes different of neutrality could have on solving Markov decision processes.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•Theory of computation → Sequential decision making; •Computing methodologies → Reinforcement learning;

Keywords

Markov Decision Process, Risk Sensitive Markov Decision Process, Exponential Utility

1. INTRODUÇÃO

Problemas oriundos do mundo real sempre foram objeto de estudo em pesquisas científicas em diferentes campos do conhecimento. Na Computação tais problemas motivam o desenvolvimento de diferentes técnicas para uma melhor modelagem e tomada de decisão, como, por exemplo, a técnica de Aprendizado por Reforço (em inglês *Reinforcement Learning* – RL). RL é um tipo de aprendizado de máquina muito utilizado por pesquisadores por ser não supervisionado, i.e., não necessita de qualquer conhecimento *a priori* do ambiente que se quer explorar. Neste modelo de Aprendizado por Reforço, consegue-se aprender incrementalmente ao longo das iterações com o ambiente [21].

O agente com RL possui conhecimento de qual é o estado

atual (s_i) e quais ações podem ser tomadas, mas *a priori* desconhece qual a melhor ação. A partir da estimativa descontada, do armazenamento das recompensas obtidas anteriormente (r_i) e da exploração aleatória com o ambiente, o agente passa a determinar qual a melhor ação (a_i) a ser executada em cada estado, e desenvolve ao longo das iterações, o seu aprendizado (Figura 1).

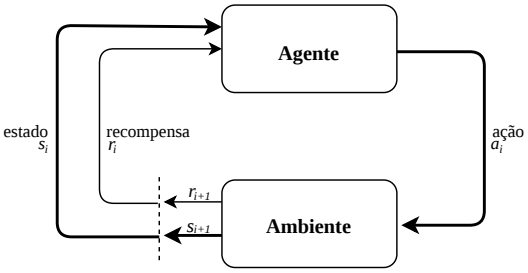


Figura 1: Ciclo do Aprendizado por Reforço [21].

A técnica RL pode obter uma solução para problemas distintos, por exemplo, (i) em simulações do jogo de futebol, em que os times se enfrentam em busca da vitória definida pela diferença de gols, (ii) na robótica móvel, em que é definida uma sequência de ações para atingir uma meta (chegar em uma sala, pegar um objeto, carregar a bateria, etc.); e (iii) no controle de ações na bolsa de valores, almejando o acúmulo financeiro.

Uma forma comum de representar problemas de RL é o Processo de Decisão Markoviano (em inglês *Markovian Decision Process* – MDP)[18], modelo que permite representar estados, ações, transições entre os estados e recompensas [21]. O risco na tomada de decisão surge a partir das incertezas associadas a eventos futuros, e é inevitável uma vez que as ações não são determinísticas, configurando um processo de decisão estocástico [20]. Assim, o risco é inerente a tomada de decisão, mas não necessariamente otimizado pelo critério do tomador de decisão, i.e., o agente. De forma que o agente modelado pode ter duas posturas distintas frente ao risco, estimá-lo ou ignorá-lo [20]. Em MDPs clássicos é considerada a tomada de decisão neutra ao risco, em que comumente o objetivo é apenas encontrar a melhor política que maximize as recompensas esperadas acumuladas [18]. O desenvolvimento de algoritmos de RL seguro, que consideram as relações de risco na tomada de decisão, é um tema pouco explorado na literatura [5]. Esses algoritmos podem ser classificados pelo critério de otimização e pelo processo de exploração usado, conforme proposto em [5].

Segundo a classificação pelo critério de otimização adotado, existem os critérios de pior caso, sensível ao risco, baseado em restrições, entre outros. Dentro de cada um desses critérios de otimização, ainda existem diferentes abordagens (Ver Figura 2). No critério sensível ao risco é incluído um fator que permite lidar com diferentes tipos de risco: propenso, neutro ou averso ao risco. Neste trabalho, estamos interessados em especial nas abordagens que utilizam a soma ponderada do retorno e risco [19, 14, 3, 7], e a utilidade exponencial [11, 2, 1, 20].

2. APRESENTAÇÃO DO PROBLEMA

Formalmente um RSMDP é uma tupla: $< \mathbb{S}$, $\mathbb{A}$, $\mathbb{T}$, $\mathbb{R}$, $\lambda >$, em que:

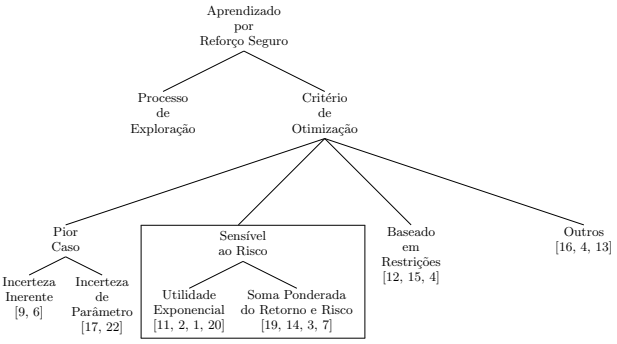


Figura 2: Abordagens de aprendizado por reforço seguro com enfoque no critério de otimização.

- $\mathbb{S}$ é o conjunto de estados do problema;
- $\mathbb{A}$ é o conjunto de ações que podem ser tomadas;
- $\mathbb{T} : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \rightarrow [0, 1]$ é uma função que define a probabilidade de transição entre estados no sistema;
- $\mathbb{R} : \mathbb{S} \times \mathbb{A} \rightarrow \mathbb{R}$ é uma função recompensa que define o custo ou recompensa em se tomar uma ação $a \in \mathbb{A}$ em um estado $s \in \mathbb{S}$; e
- λ é o fator de risco $\lambda \in (-1, 1)$.

Esta modelagem possui suporte ao risco que é implementado pelo fator λ. Se $\lambda < 0$, tem-se a propensão ao risco, em contrapartida se $\lambda > 0$, tem-se aversão ao risco, e quando o limite de $\lambda \rightarrow 0$, tem-se a neutralidade no risco, i.e. é equivalente ao MDP clássico [11].

Considerando o critério sensível ao risco, existem duas principais abordagens, que podem ser usadas: a *soma ponderada de retorno e risco*, e a *utilidade exponencial*.

A abordagem de soma ponderada do retorno e risco é uma abordagem recente na literatura de aprendizado por reforço seguro sensível ao risco, entre os trabalhos que usam essa abordagem estão [19, 14, 3, 7]. A função objetivo desta abordagem é $\max_{\pi \in \Pi}(E_\pi(R) - \lambda\omega)$, em que $E_\pi(R)$ é a expectativa de retorno da política π, λ é o parâmetro de risco e ω define o risco considerado no modelo, esse valor varia conforme a modelagem adotada pelo autor.

A principal vantagem da abordagem de soma ponderada de retorno e risco é que possibilita a variação entre as diferentes atitudes ao risco de aversão para propensão e vice-versa sem grandes prejuízos na otimização da recompensa a longo termo. Uma limitação desta abordagem é que muitas vezes uma ponderação conveniente pode não ser facilmente identificada. Outra desvantagem é que quando um critério conservador é adotado, a política resultante pode ser extremamente pessimista para um determinado problema. Nesse caso, a verdadeira utilidade das ações ao longo termo também é perdida, e o modelo ainda não é capaz de detectar as situações de risco a se evitar nos passos iniciais da execução.

A abordagem com função exponencial tem utilidade $u(R) = -\text{sgn}(\lambda)\exp(\lambda R)$ [11, 2], em que λ é o fator de risco, *sgn* denota o sinal e R é o retorno. Nesta abordagem o custo de se calcular o exponencial do retorno, i.e., $exp(\lambda R)$ pode ser proibitivo a muitas aplicações em especial quando o R é um número muito grande, neste caso ocorre o problema de computação *overflow* (estouro) [8]. Esse estouro da representação numérica alocada do dado em memória pode

comprometer o cálculo completo da exponencial. Uma implementação cuidadosa precisa atentar-se a capacidade de representação do tipo de dado utilizado e da operação para o cálculo em cada etapa, evitando ultrapassar os respectivos limites numéricos que existem tanto para valores positivos quanto negativos.

Mesmo com esse problema para o seu cálculo, a abordagem de função exponencial é considerada robusta e necessita de um aprimoramento para torna-se mais viável de ser aplicada a problemas com horizonte infinito e com um conjunto de estados maior. Assim, o desenvolvimento de estratégias para contornar tais limitações dos atuais algoritmos de aprendizado por reforço usando esta abordagem é importante e é o foco desta pesquisa de mestrado.

3. PROPOSTA DE SOLUÇÃO

Esta pesquisa é um esforço no aprimoramento de técnicas de aprendizado por reforço que usam a abordagem com função exponencial para RSMDPs.

Primeiro, para resolver o problema de *overflow*, será utilizada a operação inversa da função objetivo, isto é utilizar o logaritmo. O uso de logaritmo para RSMDPs foi proposto inicialmente por [10]. O cálculo da utilidade exponencial será exato a fim de garantir uma maior precisão na tomada de decisão. O cálculo completo da exponencial permite garantir que o algoritmo alcance de fato a convergência para uma política ótima estacionária.

Segundo, um algoritmo de aprendizado por reforço considerando a abordagem de função exponencial será proposto. O algoritmo é uma versão do algoritmo Q-Learning para RSMDPs com utilidade exponencial.

O algoritmo proposto atualiza a função $Q(s, a)$, a qualidade do par estado-ação, por meio da equação:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha u(r_t + \gamma \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t)),$$

em que $0 < \alpha \leqslant 1$ é a taxa de aprendizado, r_t é a recompensa obtida na iteração t, $0 < \gamma \leqslant 1$ é o desconto no aprendizado e u é a função de utilidade exponencial. A próxima ação é aleatoriamente escolhida conforme a política que se baseia nos atuais valores de Q. Conforme o sistema interage com o ambiente, o agente obtém as respectivas recompensas e se move para o próximo estado s'. O processo continua até que satisfaça uma condição de parada que pode ser atingir um estado meta ou realizar uma quantidade de iterações máxima.

4. AVALIAÇÃO DA SOLUÇÃO

Será feita a comparação do desempenho dos algoritmos de aprendizado por reforço clássicos, Q-Learning [23] e Sarsa [21] e os algoritmos sensíveis ao risco de [20] e [14] com o novo algoritmo proposto. Os aspectos de maior interesse a serem avaliados são: (i) o tempo de convergência do aprendizado pois é almejado um algoritmo que convirja mais rapidamente; e (ii) a política obtida, a fim de verificar se realmente o algoritmo consegue retornar uma política aversa ou propensa ao risco em cada um dos domínios de teste.

Para uniformizar os agentes a serem avaliados, todos os algoritmos serão baseados na abordagem de exploração ϵ-greedy. Uniformizar a forma de exploração dos algoritmos é importante para realizar uma comparação com menos viés da interferência do processo exploratório no aprendizado do algoritmo. Além disso, serão usados os mesmos valores para o fator de desconto γ, o fator de aprendizado α, a taxa de exploração ϵ e o fator de risco λ para todos os algoritmos.

Para compreender o papel do fator do risco na qualidade do aprendizado serão avaliados diferentes valores desse fator. Para propensão serão usados $\lambda \in \{-0,99; -0,9; -0,7; -0,5; -0,3; -0,1\}$, para neutralidade $\lambda = 0$, e para aversão ao risco $\lambda \in \{0,99; 0,9; 0,7; 0,5; 0,3; 0,1\}$.

Foram escolhidos dois domínios de teste para realizar os experimentos: o problema de travessia do rio e o simulador de futebol de duas dimensões, que são descritos a seguir.

4.1 Travessia do rio

O problema consiste na travessia do agente em um *grid* ($Nx \times Ny$) do canto inferior esquerdo para o respectivo canto inferior direito, que é o estado meta para o problema. Nesse domínio existe apenas um agente com 4 ações de movimento (Norte, Sul, Leste, Oeste). A travessia só pode ser realizada (i) nadando no rio ou (ii) caminhando pela borda do rio até chegar a uma ponte. Porém, o rio leva a uma cachoeira em que o agente pode cair ou até morrer.

4.2 Simulador de futebol

No simulador de futebol, existem duas equipes em um *grid* de duas dimensões. Neste domínio o objetivo é alcançar um maior saldo de gols para o seu time. As ações permitidas são de movimento, chute, retomada de bola e de parada. O estado atual do problema é descrito pela posição de todos os jogadores mais a bola no *grid*. O problema tem uma quantidade de estados relevante conforme o tamanho do *grid* adotado e a quantidade de jogadores por time, o que implica em um alto custo computacional.

5. ATIVIDADES REALIZADAS

Foram implementados os algoritmos de aprendizado por reforço clássicos Q-Learning [23] e Sarsa [21] e adaptados para a sensibilidade ao risco utilizando a abordagem de soma ponderada do retorno e risco de [14]. Atualmente está em desenvolvimento o novo algoritmo que se baseia em utilidade exponencial e a implementação de [20].

Os dois domínios de teste escolhidos já foram implementados e possuem uma interface para facilitar a inclusão de novos algoritmos.

Os resultados preliminares indicam que em geral pode se obter um curva com uma maior recompensa acumulada utilizando a neutralidade ao risco em menor quantidade de iterações. Todavia uma atitude distinta a neutra, pode obter uma convergência melhor a depender do experimento. Nos experimentos a aversão ao risco trouxe um ganho maior quando executado em um ambiente sem oponentes e a propensão ao risco por sua vez se mostrou mais efetiva quando existia um oponente.

6. CONCLUSÃO

A área de pesquisa de aprendizado por reforço sensível ao risco possue alta aplicabilidade para resolver diferentes problemas e propõe uma abordagem mais sofisticada para a modelagem de variáveis que por muito tempo foram ignoradas no contexto da tomada de decisão.

O aprimoramento de algoritmos de aprendizado por reforço para RSMDPs que usam uma função exponencial é de grande importância. Isso é necessário para consolidar um modelo mais independente e livre de domínio, que seja possível de ser aplicado a problemas com diferentes tamanhos

de conjuntos de estados e ações e sem grandes impactos na sua complexidade.

Este projeto de mestrado tem como objetivo propor uma nova estratégia de aprendizado por reforço sensível ao risco, que utilize a abordagem de utilidade exponencial de maneria mais eficiente. O que possibilitará que esta abordagem tenha um menor custo computacional e consiga inclusive resolver problemas com um conjunto maior de estados e ações.

7. AGRADECIMENTOS

Os autores agradecem à Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro recebido para as atividades de pesquisa que incluem este projeto.

8. REFERÊNCIAS

[1] A. Basu, T. Bhattacharyya, and V. S. Borkar. A learning algorithm for risk-sensitive cost. *Mathematics of Operations Research*, 33(4):880–898, 2008.

[2] V. S. Borkar. A sensitivity formula for risk-sensitive cost and the actor–critic algorithm. *Systems & Control Letters*, 44(5):339–346, 2001.

[3] P. Campos and T. Langlois. Abalearn: Efficient self-play learning of the game abalone. *INESC-ID, neural networks and signal processing group*, 2003.

[4] D. D. Castro, A. Tamar, and S. Mannor. Policy gradients with variance related risk criteria. In *29th ICML*, pages 935–942, 2012.

[5] J. García and F. Fernández. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.*, 16(1):1437–1480, Jan. 2015.

[6] C. Gaskett. Reinforcement learning under circumstances beyond its control. In *CIMCA*, 2003.

[7] P. Geibel and F. Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *J. Artif. Intell. Res.*, 24:81–108, 2005.

[8] A. Gosavi. Reinforcement learning: A tutorial survey and recent advances. *INFORMS J. on Computing*, 21(2):178–192, Apr. 2009.

[9] M. Heger. Consideration of risk in reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 105–111, 1994.

[10] D. Hernández-Hernández and S. I. Marcus. Risk sensitive control of markov processes in countable state space. *Syst. Control Lett.*, 29(3):147–155, Nov. 1996.

[11] R. A. Howard and J. E. Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

[12] Y. Kadota, M. Kurano, and M. Yasuda. Discounted markov decision processes with utility constraints. *Comput. Math. Appl.*, 51(2):279–284, Jan. 2006.

[13] D. G. Luenberger. *Investment science*. Oxford University Press, 2 edition, 2013.

[14] O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49(2):267–290, 2002.

[15] T. M. Moldovan and P. Abbeel. Safe exploration in markov decision processes. *CoRR*, abs/1205.4810, 2012.

[16] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 799–806, 2010.

[17] A. Nilim and L. El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Oper. Res.*, 53(5):780–798, Sept. 2005.

[18] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[19] M. Sato, H. Kimura, and S. Kobayashi. Td algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 16(3):353–362, 2001.

[20] Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer. Risk-sensitive reinforcement learning. *Neural computation*, 26(7):1298–1328, 2014.

[21] R. Sutton and A. Barto. *Reinforcement learning: An introduction*, volume 116. Cambridge Univ Press, 1998.

[22] A. Tamar, H. Xu, and S. Mannor. Scaling up robust mdps by reinforcement learning. *CoRR*, abs/1306.6189, 2013.

[23] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Seleção de atributos para mineração de processos na gestão de incidentes

Alternative Title: Attribute selection for process mining on incident management

Claudio A. L. do Amaral
Universidade de São Paulo
03828-000, São Paulo - SP
claudio.amaral@usp.br

Sarajane M. Peres
Universidade de São Paulo
03828-000, São Paulo - SP
sarajane@usp.br

RESUMO

O processo de tratamento de incidentes é o mais adotado pelas empresas, porém, ainda carece de técnicas que possam gerar estimativas assertivas para o tempo de conclusão. Este trabalho apresenta uma abordagem para uso em mineração de processos, capaz de descobrir o modelo do processo sob a forma de um modelo de transição de estados anotado e propor meios automatizados de escolha dos atributos que o descrevam adequadamente e possam gerar estimativas realistas sobre o tempo necessário para resolução. A estratégia resultante da aplicação de técnicas de seleção de atributos - filtro e invólucro - deverá permitir a geração de modelos mais precisos e com algum grau (desejado) de generalização. A solução proposta neste trabalho deve representar uma melhoria na mineração de processos, no contexto da criação de modelos de transição de estados anotados e seu uso como gerador de estatísticas para o processo nele modelado.

Palavras-Chave

Mineração de processos, Gestão de incidentes, Framewok ITIL, Seleção de atributos, Filtro e invólucro.

ABSTRACT

The incident management process is widely adopted by companies. However, still lacks techniques that can generate precise estimates for the completion time. This work presents an approach to be used in process mining that is able to find out the real process model as an annotated transition system and proposes automated means for selecting attributes that describe it accordingly, in order to generate realistic estimates of time to resolution. The resulting strategy of application feature selection techniques - filter and wrapper - should be able to generate more accurate models with some degree (desired) of generalization. The solution proposed in this paper should be an improvement in process mining on the context of annotated transition system creation and its use as a statistics generator for the modeled process.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•**Computing methodologies** → **Feature selection;**

Keywords

Process mining, Incident management, ITIL framework, Attribute selection, Filter and wrapper.

1. INTRODUÇÃO

A melhoria da eficiência e eficácia em áreas operacionais são metas almejadas em todas as organizações. Os cenários são complexos e por vezes objetivos antagônicos precisam ser atingidos. Dessa forma, é preciso realizar as tarefas que produzam os resultados necessários da melhor maneira possível. Alguns exemplos são a otimização de recursos, a redução de custos, a melhoria da satisfação dos clientes etc. Nesse contexto, a utilização de ferramentas de análise de dados e processos, surge como forma de apoiar as decisões e tornar tais cenários mais previsíveis. Em alguns setores, como o de serviços na área de operações de processos de tecnologia, essa busca tem difundido a utilização de diversos "modelos de boas práticas". O mais utilizado é o *Information Technology Infrastructure Library* (ITIL) [4].

No "*framework* ITIL", destaca-se como mais utilizado, o processo de tratamento de incidentes [6], o qual trata das ações necessárias para corrigir falhas ou degradações. Esse processo tem a característica de gerar resultados tangíveis em um horizonte de curto prazo; e sobressai-se a contribuição na identificação de prioridades, redução do tempo de atendimento, melhoria na previsão de utilização dos recursos, entre outras formas de otimização. A formalização do processo de tratamento de incidentes tem um fator crítico, a definição de indicadores. Um dos principais é o tempo alvo para a resolução do incidente. Porém, um dificultador é a complexidade em realizar estimativas confiáveis acerca do tempo necessário para executar e completar uma instância do processo (caso). Essa lacuna está relacionada às características do próprio caso e à forma de atuação adotada por pessoas e equipes durante o tratamento dos incidentes. Geralmente, os modelos estabelecidos seguem as recomendações do ITIL. Porém, o modelo de processo real diverge do modelo formal e os indicadores que derivam de uma análise do processo, costumam apresentar informações superficiais sobre os incidentes e seus tempos de resolução.

Atualmente há um grande número de empresas que utilizam sistemas orientados a processos no suporte às suas operações. Esses sistemas registram logs de execução (log de

eventos) com informações sobre as atividades executadas. A informação disponível permite uma análise detalhada do processo, feita por meio de um procedimento de mineração de processos capaz descobrir os modelos associados [8].

Os processos podem ser de três do tipos: estruturado (lasanha), semi-estruturado e não estruturado (espaguete), como definido por Aalst [8]. O processo de incidentes é classificado como semi-estruturado, pois a maioria (mais de 80%) dos casos são tratados de maneira conhecida. Entretanto, algumas atividades requerem interpretação, podem sofrer desvios e serem guiados pela experiência. Esse cenário viabiliza a análise de suporte operacional proposta nesse trabalho. Em mineração de processos, destacam-se três atividades principais: **descoberta** de modelos, com fluxos descrevendo o processo em execução; a avaliação da **conformidade** de um determinado *trace* (conjunto de eventos referente a uma instância de processo) no log em relação a um modelo pré-determinado do processo; e a **melhoria** do processo. Em processos de tecnologia os estudos concentram-se nas atividades de descoberta [1] [2].

Há dois tipos de análises realizadas em mineração de processos. A primeira utiliza de registros do log de eventos dos casos concluídos (*post mortem*). A segunda análise utiliza os registros de log parciais de casos em execução, ou seja, não concluídos. Os dados do primeiro tipo de análise serão utilizados para criar o modelo preditivo e os dados da segunda análise serão utilizados no processo de predição.

Na literatura, há iniciativas utilizando técnicas de predição em mineração de processos com o objetivo de implementar melhorias. Porém, tais iniciativas não têm mostrado preocupação com a fase do pré-processamento de dados, frequentemente usada na área de mineração de dados. Uma das fases de pré-processamento, a seleção de atributos, permite melhorar o desempenho de predição. Dentre as técnicas, os métodos de filtro [3] atuam de forma independente da escolha do preditor, ou seja, um critério de classificação é definido para os atributos individuais de forma independente dos demais. Os métodos de correlação pertencem a essa categoria. A técnica de invólucro (do inglês *wrapper* [5]) usa o desempenho de predição de uma máquina de aprendizagem e técnicas de busca para avaliar a performance e selecionar os subconjuntos de atributos.

Este trabalho pretende atuar no estudo de um processo real de tratamento de incidentes e propor meios automatizados de seleção dos atributos que o descrevam, de modo que estimativas realistas sobre o tempo necessário para resolução de um incidente sejam geradas.

2. APRESENTAÇÃO DO PROBLEMA

Quando um incidente é identificado e informado pelo usuário, a expectativa principal é saber qual o tempo necessário para o restabelecimento do serviço. A estimativa apresentada, que costuma seguir o direcionamento do *framework*, é baseada em alguns atributos como: abrangência do impacto causado e a urgência informada. Tal cenário costuma ser impreciso, pois um incidente é descrito por um conjunto maior de atributos relacionados. Além disso, no transcorrer do tratamento, outros atributos são agregados. Dessa forma, o número total de atributos que descrevem o incidente é elevado (próximo de uma centena) e os cenários são relacionados a situações distintas. Diante desse contexto complexo, há um erro de estimação associado ao tempo previsto para a resolução do incidente. Existe um problema em aberto, caracterizado pela necessidade do estabelecimento de estimativas mais assertivas.

Os atributos citados descrevem as instâncias de processo, são armazenados nos sistemas de gerenciamento de incidentes e representam o seu estado atual. Além dessa informação, os sistemas possuem um log de auditoria que armazena registros referentes às diversas atividades executadas durante o tratamento do incidente. Exemplos dessas atualizações são informações referentes à interação com o solicitante, o agente que realizou o atendimento, etc. A combinação dos atributos que descrevem o incidente com informações sequenciais do log, permitem uma análise detalhada do processo e a realização de estimativas a cada evento.

A análise mencionada, utilizando a representação do processo na forma de um modelo de transição de estados anotado (MTA) [7], gera informações sobre tempos de execução e expectativas de conclusão baseada nas estatísticas agregadas ao modelo. O log resultante da combinação acima proposta, tanto em número de registros (eventos) quanto em número de atributos, possui uma granularidade que dificulta a análise e produz modelos de processos que restringem as generalizações importantes na geração de estimativas assertivas. A dificuldade está na escolha dos atributos e perspectivas de avaliação a serem utilizadas na construção do MTA para obtenção de previsões otimizadas.

As situações descritas caracterizam o cenário do problema em estudo neste trabalho, ou seja, a necessidade de utilização de métodos mais eficientes para análise do processo de tratamento de incidentes e a realização de avaliações que considerem o processo real e possam indentificar os atributos (e combinações) que influenciam seu tempo de execução.

3. PROPOSTA DE SOLUÇÃO

Esse trabalho usa a proposta de criação de um MTA para realizar estimativas de conclusão [7]. Na criação do MTA utiliza-se o *trace* de casos concluídos agrupados por instância. A estimativa do tempo para conclusão de um caso é feita com o *trace* parcial, usado para identificação do estado atual no MTA. Utiliza-se então a informação obtida de outras instâncias que passaram por esse estado para realizar a estimativa do tempo de conclusão baseado em estatísticas.

O diferencial da proposta é a aplicação de estratégias de seleção de atributos de modo que seja gerada uma lista de atributos descritivos de um incidente próxima da lista ótima. Essa lista deve viabilizar a construção de um MTA do qual estimativas assertivas podem ser obtidas. A seleção de atributos está baseada nos estudos e recomendações apresentados nos trabalhos de Guyon e Elisseef [3] e Kohavi e John [5].

No processo de seleção, serão utilizadas duas técnicas - filtro e invólucro (do inglês *wrapper*). A estratégia inicia-se pela aplicação de uma técnica de identificação de atributos redundantes, com a utilização de avaliação por coeficiente de correlação de Pearson. Nessa avaliação, os atributos redundantes são identificados e podem ser removidos. A primeira técnica, filtro, faz a seleção de atributos por meio de ordenação. Para estabelecer a ordem serão geradas avaliações para cada atributo, por meio do uso do ganho de informação e esta ordem estabelecerá quais atributos serão utilizados na construção do MTA. Pretende-se ainda, utilizar os modelos de abstração para representação dos estados: sequencia (considera a ordem dos eventos na construção do estado), conjuntos (ignora a ordem) e multi-conjuntos (considera o número de ocorrências, ignorando a ordem) [7]. Além dos

modelos de abstração, a construção do MTA utiliza o conceito de horizonte, ou seja, o número de eventos a ser utilizado na definição do estado atual. O mais utilizado é o horizonte completo, também chamado de infinito, porém, há um efeito colateral dessa utilização que é a granularidade demasiada que faz o número de estados crescer significativamente, sobretudo no modelo de abstração sequência. Os MTA gerados serão utilizados para realização de predições dos tempos de conclusão e sua assertividade poderá ser avaliada.

A segunda estratégia, invólucro, consiste em utilizar o MTA como forma de avaliação durante a seleção de atributos em um conceito de caixa-preta. Dessa forma, a cada etapa, os atributos selecionados são utilizados para gerar o modelo. Os resultados obtidos são avaliados e essa avaliação é usada no processo de busca para avaliação da precisão e tomada de decisão em relação aos próximos passos da seleção. Há duas formas de seleção por invólucro, o *forward selection*, no qual o conjunto de atributos inicia-se vazio e cresce gradualmente de acordo com a heurística de busca selecionada e o *backward elimination*, que inicia com todos os atributos e gradualmente retira os atributos caso não contribuam para a melhora de precisão do modelo de predição.

4. AVALIAÇÃO

O MTA [7] possui, em cada um de seus estados, medidas relacionadas ao tempo gasto, tempo restante e tempo de permanencia de cada instância do processo. Essas informações permitem a geração de medidas estatísticas (média, desvio-padrão, entre outros) e testes que são utilizados como forma de avaliação. A avaliação do processo de seleção de atributos será feita a partir de resultados de predições obtidas sob uma estratégia de validação cruzada, usando períodos de tempo referentes à execução de processos como *folds*, e aplicando o erro quadrático médio e a raiz do erro quadrático médio sobre predições referentes ao período usado como *folds* de teste. Pretende-se também realizar a avaliação do grau de conformidade do modelo gerado contra os dados de validação, ou seja, avaliar se há instâncias de processo que não são possíveis de reprodução no modelo criado.

Os trabalhos são desenvolvidos utilizando-se um conjunto de dados real, obtido de uma solução de gerenciamento de incidentes da plataforma *ServiceNow*^{TM}. Esse conjunto de dados que possui um total de vinte e quatro mil, novecentos e oitenta e nove instâncias de processo (24989) e cento e cinquenta e três mil quatrocentos e nove (153409) eventos e noventa e um atributos (91) por evento, será utilizado para construção dos MTA. Tais modelos serão construidos sobre três estratégias distintas: na primeira, a seleção de atributos será realizada usando o conhecimento do especialista; na segunda, serão usadas técnicas de filtro para a decisão dos atributos a serem utilizados; na terceira, serão utilizadas técnicas de invólucro. Para cada etapa serão realizados experimentos de construção do MTA e teste do modelo como preditor a partir da validação cruzada, para que os resultados possam ser analisados. As análises utilizarão também informações oriundas do trabalho de Aalst [7] como referência para o *benchmark*.

5. ATIVIDADES REALIZADAS

A primeira atividade realizada foi a exploração do modelo de processo associado ao processo real de gestão de incidentes advindo dos dados da plataforma *ServiceNow*^{TM}. Tal análise foi realizada com as ferramentas ProM [1] e Disco [2].

Já esta finalizada a experimentação com o primeiro tipo de conjunto de dados que utiliza o conhecimento do especialista para seleção de atributos. Nesse cenário, duas abordagens foram avaliadas, com um e dois atributos respectivamente. O resultado obtido apontou que, no primeiro caso, o desvio padrão chegou a ter o valor de duas vezes a média, ao passo que, ao utilizar dois atributos esse valor foi reduzido para metade da média, ou seja, uma diferença significativa na precisão. Porém, um efeito colateral foi o aumento do número de estados distintos do modelo, que eram sessenta e três (63) e passaram a duzentos e sessenta e nove (269). Esses resultados confirmam a hipótese que a utilização de atributos especificos pode resultar em mais precisão ao modelo, porém, há também um indicativo que o modelo pode ter uma redução na sua capacidade de generalização. Houve uma dificuldade em utilização da ferramenta ProM, devido a limitações do plug-in de geração do MTA em tratar o log de eventos de incidentes. Por esse motivo, foi desenvolvida uma implementação simplificada, em termos de interface gráfica, que possibilita realizar os experimentos necessários de forma mais escalável.

Atualmente estão sendo conduzidos os experimentos com o segundo tipo de conjunto de dados, que são a seleção de atributos feita por meio da técnica de filtro com os modelos de abstração citados e variação do horizonte. Há um indicativo que o horizonte tem relevância na precisão do modelo e sua capacidade de generalização. Essa avaliação é medida pelo índice percentual de não aderencia das instâncias de processo dos conjuntos de dados de testes e validação.

A próxima etapa é a realização dos experimentos com a técnica de invólucro para que os resultados possam ser avaliados.

6. CONCLUSÃO

A solução proposta neste trabalho deve representar uma melhoria na mineração de processos, no contexto específico da criação de MTA e no seu uso como um gerador de estatísticas mais precisas para o processo nele modelado.

A estratégia resultante da aplicação de técnicas de seleção de atributos deverá ser capaz de propiciar a geração de MTA mais precisos, com algum grau (desejado) de generalização. Por consequencia, as estimativas de tempo para conclusão, relacionadas ao processo de gestão de incidentes, devem ser mais precisas.

Outro fator relevente é a aplicabilidade da abordagem proposta, pois o processo de incidentes é amplamente utilizado nas organizações [6]. Assim, há possibilidade da tecnica desenvolvida ser integrada a produtos de software na área de gestão de processos.

7. REFERÊNCIAS

[1] A. Bautista, S. Akbar, A. Alvarez, T. Metzger, and M. Reaves. Process mining in information technology incident management: A case study at Volvo Belgium. In *Proceedings of 3rd BPI Challenge*, 2013.

[2] E. Dudok and P. van den Brand. Mining an incident management process. In *Proceedings of 3rd BPI Challenge*, 2013.

[1] ProM http://www.promtools.org/

[2] Disco https://fluxicon.com/disco/

[3] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[4] itSMF International. itsmf 2013 global survey on IT service management, 2013.

[5] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 1-2(97):273–324, 1997.

[6] M. Marrone, F. Gacenga, A. Cater-Steel, and L. Kolbe. IT service management: A cross-national study of ITIL adoption. *Communic. of the Assoc. for Inf. Syst.*, 34(1):865–892, 2014.

[7] W. van der Aalst, M. Schonenberg, and M. Songa. Time prediction based on process mining. *Information Systems*, 36(2):450–475, 2011.

[8] W. M. P. van der Aalst. *Process Mining - Discovery, Conformance and Enhancement of Business Processes*. Springer, 1st edition, 2011.

Arquitetura para Sistemas de Recomendação de Notícias: Uma Abordagem Híbrida e Baseada em Casos

Alternative Title: Architecture for News Recommendation Systems: A Hybrid and Case-Based Approach

José Luiz M. Pagnossim
Universidade de São Paulo, EACH
Rua Arlindo Béttio, 1000
Sao Paulo – SP, Brasil
+55 (11) 98058-7052
pagnossim@usp.br

Sarajane Marques Peres
Universidade de São Paulo, EACH
Rua Arlindo Béttio, 1000
Sao Paulo – SP, Brasil
+55 (11) 3091-8897
sarajane@usp.br

RESUMO

Sistemas de Recomendação (*SR*) são capazes de sugerir itens por meio de similaridade e do histórico de interações entre usuários e sistema. Para que os *SR* atendam às expectativas dos usuários, estes devem ser eficazes em: modelar os dados; recuperar a informação; combinar similaridade e relevância; aplicar modelos preditivos ou descritivos; e, finalmente, evoluir a inteligência do sistema. *SR* podem apresentar limitações relacionadas à falta de dados históricos e ausência de indicadores de popularidade dos itens, podem também incorrer em problemas que resultem em recomendações: aleatórias, previsíveis, inadequadas, irrelevantes e incapazes de surpreender o leitor. Para enfrentar estes desafios, este artigo propõe a definição de uma arquitetura híbrida e baseada em casos para recomendação de notícias que seja capaz de melhorar indicadores de aceite, popularidade e serendipidade das recomendações, por meio de sugestões de notícias similares, relevantes, diversificadas, que despertem surpresa ao leitor e considerem o perfil do usuário.

Palavras-chave

Sistemas de recomendação; Recomendação de notícias; Arquitetura de sistemas; Raciocínio baseado em casos; Recomendação baseada em casos.

ABSTRACT

Recommendation Systems (RS) can suggest items using similarity and the history of interactions between users and system. To meet user's expectations, RS should be effective in: data modeling; information retrieval; combination of similarity and relevance; application of predictive or descriptive models; and finally, in the system intelligence evolution. RS may present limitations related to the lack of historical data and lack of items popularity indicators, may also have problems resulting in recommendations: random, predictable, inadequate, irrelevant and unable to surprise the user. To address these challenges, this paper proposes the definition of a hybrid and case-based news recommendation architecture that can improve indicators of acceptance, popularity, and serendipity of recommendations through similar, relevant, diversified, surprise the reader and consider the user's profile.

SBSI 2017, June 5th–8th, 2017, Lavras, Minas Gerais, Brazil.

CCS Concepts

• Information Systems Applications → Miscellaneous • Other Architecture Styles → Heterogeneous (hybrid) systems.

Keywords

Recommendation systems; News recommendation; Systems architecture; Case-based reasoning; Case-based recommendation.

1. INTRODUÇÃO

Sistemas de Recomendação (*SR*) são softwares e ferramentas capazes de sugerir itens que são mais prováveis de interesse para um determinado usuário [8]. São também capazes de prover recomendações com base no histórico de interações entre usuários e sistema, entre usuários e itens, ou ainda com base em similaridade. Dentre os tipos mais comuns de *SR* estão: por filtro colaborativo; baseados em conteúdo; e, baseados em conhecimento. Os *SR* por filtro colaborativo permitem recomendar ao usuário, itens que outros usuários similares a ele gostaram no passado [1]. Os *SR* baseados em conteúdo, consideram a descrição dos itens para encontrar itens semelhantes, de forma a fazer recomendações com base na similaridade entre itens [1], normalmente os atributos para descrever um item são extraídos a partir de metadados associado ao item ou à características textuais extraídas da descrição do item [4]. Já os *SR* baseados em conhecimento criam as recomendações com base em informações de domínio e de preferências dos usuários [8]. Desde sua criação, a área de *SR* expandiu rapidamente com aplicações que recomendam filmes, páginas web, notícias, tratamentos médicos, músicas e outros produtos [7]. A implementação de *SR* envolve tarefas complexas, que podem ser resolvidas a partir da visão de áreas como Inteligência Artificial na qual encontram-se estudos referentes à metodologia de Raciocínio Baseado em Casos (*RBC*) que trabalha com o princípio que se algo funcionou no passado em uma determinada situação, muito provavelmente pode funcionar novamente em uma nova situação similar àquela passada [6]. Seguindo esse pressuposto, Smyth [9] propõe uma estratégia para uso de *RBC* em *SR*, apresentando a Recomendação Baseada em Casos (*Recomendação BC*), que consiste em um tipo particular de recomendação baseada em conteúdo que agrega a capacidade de tratar itens de forma estruturada, representados por atributos e seus respectivos valores (modelo de "caso"). Esse tipo de recomendação se aplica adequadamente ao domínio de notícias, pois pode fazer uso de informações textuais para encontrar a similaridade entre as

notícias, pode utilizar informações estruturadas como o canal de leitura da notícia e ainda armazenar as recomendações fornecidas no passado para solucionar novos problemas. A escolha de uma abordagem em *SR* pode beneficiar um aspecto da recomendação em detrimento de outro. Para minimizar esse desequilíbrio adota-se aspectos híbridos, como defendido por [3] que mescla *RBC* e Mineração de Dados. O aspecto híbrido deste trabalho, considera: tirar proveito dos diferentes tipos de *SR*; usar diferentes técnicas de resolução de problemas; e combinar a informação advinda de diferentes fontes para compor uma métrica única de recomendação. Há diferentes formas para se avaliar um sistema de recomendação e os principais métodos, segundo [5], são: experimentos *offline*; estudos do usuário; e avaliação *online*. Uma forma de avaliação *online* nos *SR* é medir se os itens sugeridos foram consumidos pelos usuários. No caso de notícias, se o usuário leu uma notícia, o consumo deste item pode ser contabilizado. Diante desse panorama, este artigo apresenta uma arquitetura para sistemas de recomendação de notícias que visa resolver problemas conhecidos em sistemas de recomendação a partir de uma abordagem híbrida, sugerindo notícias mais interessantes ao usuário.

2. APRESENTAÇÃO DO PROBLEMA

Cada um dos tipos de *SR* apresenta características que limitam o atendimento em relação à expectativa do usuário, incorrendo em problemas como *cold-start* (de item ou de usuário), que apresentam situações em que o recomendador é incapaz de fornecer recomendações significativas devido a uma falta inicial de avaliações de usuários [2]. Outros problemas, que podem ser detectados por meio da navegação em portais de notícias disponíveis no estado da prática estão relacionados à recomendações aleatórias, previsíveis, inadequadas, irrelevantes e incapazes de surpreender o usuário. Recomendar atendendo adequadamente às expectativas do usuário sob diferentes perspectivas é um problema difícil de ser superado e que merece ser mais aprofundado. Um outro desafio em *SR* está associado à forma de avaliação dos resultados. Diante deste contexto, o problema tratado por esta pesquisa está relacionado ao cenário de *cold-start*, falta de relevância, novidade, diversidade e serendipidade nas recomendações, e explora também o desafio da avaliação em *SR*.

3. PROPOSTA DE SOLUÇÃO

Para resolver os problemas citados, os *SR* devem apresentar soluções eficientes e eficazes para: modelagem dos dados que suportarão a predição da recomendação; recuperação da informação inerente a todos os atributos que descrevem os dados; combinação dessa informação dentro de métricas de similaridade, relevância ou adequabilidade; criação de modelos preditivos ou descritivos para elaboração da recomendação; e implementação de mecanismos de evolução da inteligência do sistema de forma que ele seja capaz de aprender a partir da interação com o usuário. Para isso, os componentes computacionais dos *SR* devem ser organizados de forma a constituir uma arquitetura de recomendação, que neste artigo, é definida como uma arquitetura híbrida para sistemas de recomendação de notícias apoiada em uma abordagem baseada em casos. O desenho dessa arquitetura é demonstrado na figura 1.

3.1 Módulos da Arquitetura

A seguir são especificados os módulos da arquitetura de recomendação proposta por este artigo: Fonte de Dados: Trata-se do portal de notícias na internet que serve de fonte para montagem da base de notícias; Captura de Dados: Obtém o conteúdo (formato bruto) da notícia a partir do portal na internet; Transformação de Dados: Interpreta os dados brutos capturados do portal, armazena os metadados da notícia na Base de Casos e disponibiliza o conteúdo em formato texto para ser processado; Texto: Conteúdo das notícias disponibilizado em arquivos texto; Mineração de Dados: Pré-processa os dados textuais das notícias para gerar representações vetoriais, calcular a similaridade e gerar agrupamentos entre as notícias; Base de Casos: Repositório responsável por modelar e armazenar as notícias em formato de "caso"; CRUD: Módulo responsável pelo acesso e manipulação dos dados apresentados em tela. Acrônimo que em inglês significa Create, Read, Update e Delete; Recomendador:Núcleo de recomendação da arquitetura, define critérios e estratégias de recomendação; Apresentação: Camada que faz a interface entre o usuário e a arquitetura de recomendação de notícias.

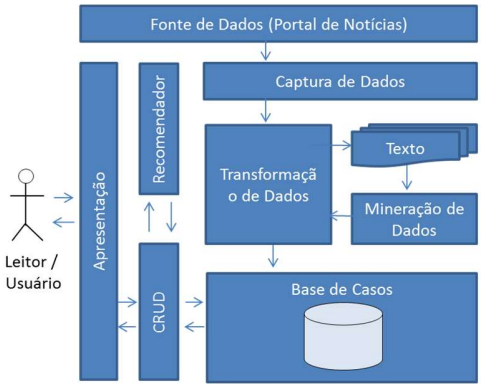


Figura 1. Arquitetura de recomendação de notícias

3.2 Aspecto híbrido da solução

O aspecto híbrido da solução, considera tratar *SR* por filtro colaborativo, baseado em conteúdo e baseado em conhecimento, usar as técnicas de resolução de problemas como *RBC* e mineração de dados e combinar a informação proveniente de diferentes fontes de decisão para compor uma métrica única a ser usada para gerar a recomendação.

3.3 Critérios para recuperação da informação

A recomendação é gerada por meio de uma função que considera diferentes critérios de recuperação da informação (tabela 1). Esses critérios são combinados com estratégias de recomendação, que fazem uso de mineração de texto, interação do usuário, análise de perfis e da *Recomendação BC*.

Para padronizar os termos usados na sequência, é importante conceituar que um leitor *E* é uma pessoa que acessou o portal de notícias e navegou anonimamente pelo sistema. Já um usuário *U*, é uma pessoa que se cadastrou no sistema, fornecendo informações que possibilitarão sugestões de acordo com seu perfil. Notícia de origem (*NO*) é a notícia que está sendo lida por *E* ou por *U* e notícia de destino (*ND*) é uma notícia a ser recomendada. Canal é uma seção do portal que organiza as notícias por assuntos.

Tabela 1. Critérios para recuperação da informação

Critério	Descrição
CR_1	Recupera a notícia mais lida dentro do canal
CR_2	Recupera a notícia mais curtida do canal
CR_3	Recupera *ND* (mais recomendações aceitas)
CR_4	Recupera *ND* com maior nota de avaliação
CR_5	Recupera *ND* com melhor serendipidade

CR_6	Recupera *ND* com maior similaridade de *NO*
CR_7	Recupera a notícia mais lida do portal
CR_8	Recupera a notícia mais curtida do portal
CR_9	Recupera uma notícia aleatória (toda base)
CR_{10}	Recupera uma notícia aleatória que não tenha sido recomendada anteriormente
CR_{11}	Recupera uma notícia aleatória, não recomendada anteriormente e seja do mesmo canal de *NO*
CR_{12}	Recupera a notícia mais lida (canal favorito do *U)*

3.4 Estratégias e definição da recomendação

Na lógica da recomendação, os critérios de recuperação das notícias devem ser combinados com estratégias de recomendação visando adequar as sugestões às exigências de *E* e *U*. Tais estratégias podem ser construídas de diferentes formas, e para o contexto de apresentação da arquitetura conceitual, é pertinente discutir algumas das estratégias possíveis em termos dos problemas de recomendação que ela objetiva resolver (tabela 2).

Tabela 2. Estratégias de recomendação e problema associado

Estratégia	Critérios	Problema associado
ES_1	$CR_{6,9,11}$	*Cold-start* de *U* e critérios para *E*.
ES_2	CR_1 a CR_{11}	Previsibilidade, Novidade, Diversidade e Serendipidade.
ES_3	ES_2, CR_{12}	Personalização e preferências (*U)*.

ES_1 é usada em situações onde não há indicadores de interação entre usuário e o sistema, enquanto que a ES_2 agrega critérios que dependem de indicadores de navegação do usuário. Por fim, a ES_3 agrega um critério que considera as preferências do usuário, considerando como pré-requisito que o usuário tenha se cadastrado no sistema e se identificado na sessão de navegação.

Para implementar a métrica única de recomendação é utilizado um raciocínio de competição, em que participam as notícias selecionadas pelos critérios para recuperação da informação discutidos anteriormente. A fórmula da métrica única considera a soma das posições em que a notícia ficou na competição. Quanto menor essa soma, melhor é a posição da notícia na lista final ordenada. Uma simulação do uso da métrica híbrida é apresentada na figura 2.

	CR_1	CR_2	CR_3	CR_4	CR_5	CR_6	CR_7	CR_8	CR_9	CR_{10}	CR_{11}	CR_{12}	SRR
N_1	1	2	2	2	12	2	6	12	5	5	12	8	69
N_2	2	1	12	3	11	3	2	11	6	6	11	7	75
N_3	3	12	1	5	10	5	3	10	7	7	10	6	79
N_4	5	9	11	1	2	7	5	2	8	8	9	5	72
N_5	4	10	9	4	1	4	4	7	9	9	8	4	73
N_6	7	11	10	7	9	1	7	9	10	10	7	3	91
N_7	6	8	8	6	8	6	1	8	11	11	6	2	81
N_8	8	7	7	8	7	8	8	1	12	12	5	12	95
N_9	9	5	5	9	5	9	9	5	1	2	4	11	74
N_{10}	10	6	6	10	6	10	10	6	2	1	3	10	80
N_{11}	12	3	3	12	3	12	12	3	3	4	1	9	77
N_{12}	11	4	4	11	4	11	11	4	4	3	2	1	70

Figura 2. Simulação de cálculo da métrica única

A figura 2 traz as seguintes informações: $\underline{CR_n}$: São critérios de recuperação da informação; $\underline{N_m}$: São notícias selecionadas para recomendação; $\underline{SRR}$: System recommendation rate, representa o indicador único para recomendação que é dado pela soma de todas as posições em que a notícia ficou na competição.

Com base no indicador único da recomendação, o sistema gera uma lista de recomendação em ordem crescente de *SRR*. Se ocorrer um empate, os critérios de desempate são, em ordem de prioridade: CR_6, CR_3, CR_4, CR_1, CR_2. Se persistir o empate, a escolha da notícia vencedora utiliza o metadado referente à data/hora de publicação da notícia, priorizando a notícia mais recente e garantindo que não haja empate na recomendação final.

Adicionalmente às estratégias de recomendação apresentadas previamente, a solução faz uso da *Recomendação BC*, e a partir dos recursos desse tipo de recomendação é implementada a ES_4 para situações em que a *NO* já tenha sido anteriormente consumida, e nesta ocasião, a lógica da *Recomendação BC* armazena informações sobre tal consumo na forma de "casos" (conforme quíntupla na tabela 3). Na reincidência da leitura de uma notícia, a lista de recomendação associada ao "caso" pode ser reutilizada. A reutilização pode ser feita de forma direta, fazendo exatamente a mesma recomendação que foi feita no passado do sistema, ou pode ser adaptada (ES_5), para isso, o sistema recupera a *ND* mais similar em relação à *NO* (com base no conteúdo da notícia) e como resultado dessa estratégia, novas listas de recomendações são criadas e novos "casos" são armazenados na base de casos.

Tabela 3. Formas de armazenamento do caso

Forma do *caso*	Elementos Relacionados
Um "*caso*" gerado a partir da interação do *E*	{*E*, Interações, *NO*, *ND*, Lista de Notícias recomendadas para *E*}
Um "*caso*" gerado a partir da interação do *U*	{*U*, Interações, *NO*, *ND*, Lista de Notícias recomendadas para *U*}

4. AVALIAÇÃO DA SOLUÇÃO

Um experimento *offline* é realizado usando um conjunto de dados pré-coletados de usuários selecionando ou classificando itens [5]. Estudo do usuário, é conduzido por meio de um conjunto de cenários de teste e estimulando os usuários a executarem tarefas de interação com o sistema, enquanto isso, seus comportamentos são observados e coletados [5]. A avaliação *online* está interessada em medir o comportamento do usuário quando este interaje com diferentes sistemas de recomendação, por exemplo, se usuários de um sistema aceitam as recomendações com mais frequência em relação aos usuários de outro sistema, pode-se concluir que o primeiro sistema é superior ao segundo [5]. A experiência que fornece a evidência mais forte quanto ao verdadeiro valor do sistema é uma avaliação *online*, em que o sistema é usado por usuários que executam tarefas reais [5]. Neste trabalho, o método adotado é de avliação *online***,** medindo indicadores capturados a partir da navegação dos usuários durante três sessões de experimentação, por meio de um protótipo de portal de notícias, construído para provar o conceito da arquitetura proposta neste artigo, atuando como interface entre o sistema e os participantes do experimento. O formato do experimento é apresentado na tabela 4.

No domínio de notícias, a recomendação pode ser considerada aceita (consumida) em diferentes níveis. O primeiro mede se *E* ou *U* aceitou uma notícia recomendada pelo sistema. A partir do momento em que *E* ou *U* está lendo uma notícia, outros níveis de consumo são mensurados, por exemplo a curtida sobre uma notícia ou ainda uma nota de avaliação da recomendação na escala de um a cinco estrelas.

Tabela 4. Formato do Experimento

Sessão	Característica da recomendação	Objetivo da coleta
1	Baseada no recomendador do portal EBC[1]	Linha base para comparação do EBC[1]
2	Recomendador desta pesquisa sem indicadores de navegação	Linha base para comparação sem indicadores prévios
3	Recomendador desta pesquisa com indicadores da sessão anterior	Linha base para comparação em relação aos anteriores

O resultado da avaliação, consiste na comparação dos indicadores gerados pelos participantes durante as três sessões de experimentação, sendo que é esperado que os indicadores da segunda sessão sejam melhores que os da primeira e os indicadores da terceira sessão sejam melhores que os da segunda, em termos de notícias lidas, recomendações aceitas, curtidas, nota da recomendação e serendipidade.

5. ATIVIDADES JÁ REALIZADAS

Com o objetivo de confirmar a hipótese levantada pela pesquisa e conduzir o projeto, as seguintes atividades foram realizadas: estudos exploratórios; busca exploratória pelo estado da arte; construção de um protótipo de portal de notícias; defesa da qualificação; publicação de artigo científico apresentando a proposta da arquitetura conceitual; projeto do experimento com usuário e escrita de projeto para comitê de ética. Como parte da arquitetura de recomendação de notícias, as seguintes atividades foram realizadas: definição da fonte de dados para obtenção de notícias em língua portuguesa; obtenção de autorização de uso de notícias de um portal real (EBC[1]) definição das características e formato do *corpus* de notícias; modelagem da base de casos; desenho conceitual da arquitetura; e, a formalização do núcleo de recomendação da arquitetura. O projeto encontra-se em fase de submissão do projeto ao conselho de ética, construção dos componentes da arquitetura para prova de conceito e finalização da obtenção das notícias vindas do portal EBC[1].

6. CONCLUSÃO

O projeto em questão levantou temas relevantes para a área de *SR* como o uso da metodologia *RBC* e da Recomendação BC, e no que se refere à proposição de uma arquitetura híbrida de recomendação baseada em: filtro colaborativo; conteúdo; conhecimento; e casos. Destaca-se ainda o caráter híbrido das medidas de recuperação e avaliação de itens sob diferentes aspectos desejáveis a uma recomendação. A arquitetura conceitual, definida por este projeto, cobre algumas lacunas deixadas por muitos dos atuais *SR* de notícias do estado da prática, e evolui propostas encontradas no estado da arte, principalmente com relação ao atendimento dos aspectos relacionados à similaridade, relevância, novidade, diversidade e serendipidade. As contribuições da arquitetura proposta por este trabalho estão relacionadas a pelo menos duas áreas. A primeira é a área de comunicação, em que se enquadra um portal de conteúdo, que encontrará potencial para evoluir o processo de recomendação de notícias, vislumbrando benefícios como maior captação de leitores e usuários. A segunda é a área de desenvolvimento de sistemas, que pode reaproveitar a arquitetura, adaptando-a para outros domínios de dados e negócios.

7. REFERÊNCIAS

[1] Brunialti, L. F. et al. Aprendizado de máquina em sistemas de recomendação baseado em conteúdo textual: Uma revisão sistemática. In: Anais do XI Simpósio Brasileiro de Sistemas de Informação (SBSI), 2015. p. 203-210

[2] Brusilovsky P. et al. The adaptative Web Methods and Strategies of Web Personalization. Springer, 2007. P 311-312.

[3] CHAUDHURI, C.; CHAUDHURI, A. Detection of verbatim or partial duplication from multiple source documents using data mining techniques and case-based reasoning methodologies. Emerging Applications of Information Technology, International Conference on, IEEE Computer Society, p. 129-132, 2011.

[4] Gemmis, M; Lops, P; Musto C; Narducci F; Semeraro G: Semantics-Aware Content-Based Recommender System, In: RICCI, F. et al. (Ed.). Recommender Systems Handbook Second Edition: Springer, 2015. p. 119-159.

[5] Gunawardana A; Shani G: Evaluating Recomender Systems, In: RICCI, F. et al. (Ed.). Recommender Systems Handbook Second Edition: Springer, 2015. p. 265-308.

[6] Kolodner, J. Case-Based Reasoning. 1. ed.: Morgan Kaufmann Publishers Inc., 1993.

[7] Lops, P.; Gemmis, M. de; Semeraro, G. Content-based recommender systems: State of the art and trends. In: RICCI, F. et al. (Ed.). Recommender Systems Handbook: Springer, 2011. p. 73-105.

[8] Ricci, F. et al. Recommender Systems Handbook Second Edition: Springer, 2015.

[9] Smith, B. Case-based recommendation. In: Brusilovsky, P.; Kobsa, A.; Nejdl, W. (Ed.). The Adaptive Web: Springer, 2007. p. 342-376.

[1] www.ebc.com.br

Avaliação de estratégias heurísticas para implementação de coagrupamento aplicado a dados textuais

Alternative Title: Evaluation of heuristic strategies for implementation of coclustering applied to textual data

Alexandra K. Ramos Diaz
Universidade de São Paulo
03828-000, São Paulo, Brasil
katy.rd@usp.br

Sarajane M. Peres (Orientador)
Universidade de São Paulo
03828-000, São Paulo, Brasil
sarajane@usp.br

RESUMO

Coagrupamento é uma tarefa de mineração de dados que permite a extração de informação relevante sobre os dados e tem sido aplicada com sucesso em uma ampla variedade de domínios, incluindo aqueles envolvem dados textuais – foco de interesse desta pesquisa. No coagrupamento, os critérios de similaridade são aplicados simultaneamente às linhas e às colunas das matrizes de dados, agrupando simultaneamente os objetos e os atributos e possibilitando a criação de cogrupos. O coagrupamento de dados textuais demanda uma representação em um modelo de espaço vetorial, que comumente leva a geração de espaços de representação de alta dimensionalidade e esparsidade no qual muitos dos algoritmos de coagrupamento não obtêm boas soluções. Embora existam trabalhos propondo otimizações, heurísticas ou não, para lidar adequadamente com tal complexidade, não foi encontrado ainda, no âmbito deste trabalho, um estudo que tivesse como objetivo específico explorar e comparar como técnicas de diferentes naturezas lidam com o problema da esparsidade e de alta dimensionalidade no problema de coagrupamento, explicando quais dificuldades encontram e como as superam. Um estudo exploratório nesse tema abre caminhos para a proposição de soluções mais eficientes e eficazes para o problema, e é o objetivo da pesquisa aqui apresentada.

Palavras-Chave

Coagrupamento, Biagrupamento, Mineração de textos, Matrizes de dados esparsas, Algoritmos heurísticos.

ABSTRACT

Coclustering is a data mining task that allows the extraction of relevant information about the data and has been applied successfully in a wide variety of domains, including those involving textual data - the focus of interest of this research. In coclustering, similarity criteria are applied simultaneously to the rows and columns of the data matrices, simultaneously grouping the objects and attributes and enabling the creation of cogroups. The coclustering of textual data demands a representation in a vector space model, which commonly leads to the generation of high dimensionality and sparsity representation spaces in which many of the coclustering algorithms do not obtain good solutions. Although there are works proposing optimizations, heuristics or not, to deal adequately with such complexity, we have not yet found, in the scope of this work, a study that had as a specific objective to explore and compare how techniques of different natures deal with the problem of sparsity and high dimensionality in the problem of coclustering, explaining what difficulties they encounter and how they overcome them. An exploratory study in this theme opens the way for proposing more efficient and effective solutions to the problem, and is the objective of the research presented here.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•**Information systems** → **Clustering and classification;** *Similarity measures;*

Keywords

Coclustering, Biclustering, Text Mining, Matrix of sparse data, Heuristic algorithms

1. INTRODUÇÃO

Descobrir conhecimento significa identificar ou receber informações relevantes, e poder processá-las e agregá-las ao conhecimento prévio de um usuário, mudando o estado de seu conhecimento atual, a fim de que determinado problema, ou situação, possa ser resolvido [5]. Neste sentido, observa-se que o processo de descoberta de conhecimento está fortemente relacionado à forma pela qual é processada a informação sobre a qual ele é gerado. Atualmente, com a disseminação das funcionalidades de produção e propagação de conteúdo online, o volume de informação disponível na forma textual é muito grande e configura-se como um ambiente propício à descoberta de conhecimento de forma automática.

Para automatizar o processo de descoberta de conhecimento no contexto de dados textuais, é necessário que os dados estejam representados de maneira apropriada para sua manipulação. A representação mais frequentemente usada

para documentos textuais é baseada no modelo de espaço vetorial [8, 17, 1, 12, 13] . Nesse modelo, cada documento é representado por um vetor, e cada coordenada desse vetor corresponde a uma dimensão descritiva de uma coleção de documentos [16]. A abordagem mais utilizada na área de mineração de dados para processamento de textos, baseada no modelo de espaço vetorial, é a representação *bag-of-words*. Nesta abordagem, cada palavra encontrada na coleção de documentos pode tornar-se uma dimensão no espaço vetorial e cada vetor que representa um documento terá um componente para cada palavra [8].

No entanto esta abordagem é caracterizada por gerar representações de alta dimensionalidade e esparsidade (grande quantidade relativa de elementos nulos ou ausentes), gerando espaços complexos que dificultam a extração da informação.

A tarefa de mineração de dados chamada "agrupamento" auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões (características comuns dos elementos) que descrevem grupos similares de dados [8, 5]. As técnicas de agrupamento operam sobre matrizes de dados (no caso de dados textuais, considerando a representação *bag-of-words*), e tem como objetivo particionar dados em subconjuntos, seguindo algum critério de similaridade entre os dados. Os subconjuntos a serem formados a partir dessas matrizes de dados contêm objetos descritos por atributos que possuem valores semelhantes entre si e diferentes dos valores dos objetos de outros subconjuntos de dados.

Contudo, as diferentes técnicas de agrupamento [2] apresentam algumas desvantagens quando os dados sob análise apresentam alta dimensionalidade e são esparsos, falhando em encontrar uma partição ótima, devido ao problema conhecido como a maldição da dimensionalidade (*Curse of dimensionality*) [15], e impactando nas análises subjetivas desejadas a partir dos resultados do algoritmo.

Uma alternativa à tarefa de agrupamento é aquela conhecida como coagrupamento (do inglês *co-clustering*), que é capaz de extrair informações diferenciadas da matriz de dados, quando comparada à informação extraída com agrupamento. Isso ocorre porque no caso do coagrupamento, os critérios de similaridade são aplicados simultaneamente às linhas e às colunas das matrizes de dados, agrupando simultaneamente os objetos e os atributos [11].

1.1 Definição de coagrupamento

Embora a tarefa de coagrupamento tenha uma definição levemente diferente de uma tarefa conhecida como biagrupamento (do inglês *bi-clustering*), os algoritmos para resolvê-las transitam entre as duas linhas de estudo. Existem diversas maneiras de visualizar o problema de coagrupamento, sendo que uma delas é a visão de encontrar submatrizes de uma matriz de dados [11]. Uma ilustração do problema de coagrupamento é apresentado na figura 1. Na parte esquerda desta figura é mostrado um conjunto de dados de dimensão 4×5 que possui as linhas (dados): Student A, Student B, Student C e Student D; e as colunas (características): Science, English, Math, Chinese e Social. Já na parte da direita desta mesma figura estão ilustrados os cogrupos hipotéticos formados pelas similaridades parciais entre os dados. Um exemplo de coagrupamento é aquela em que o primeiro grupo, de cor roxa, que representa os estudantes com bom desempenho nas disciplinas de Science e Math; o segundo grupo, de cor vermelha, é formado pelos estudantes com bom desempenho nas disciplinas de English, Chinese e Social.

Embora o coagrupamento traga uma maior flexibilidade para a análise de dados, sua implementação envolve um problema computacional complexo. Tal complexidade é estudada no âmbito dos trabalhos que propõem otimizações, heurísticas ou não para o problema de coagrupamento, mas esses trabalhos não apresentam estudos que envolvam técnicas de diferentes naturezas e se limitam a realizar explorações comparativas de seus algoritmos com outros de mesma classe. Até o presente momento, não foi encontrado ainda, no âmbito deste trabalho, um estudo que tivesse como objetivo específico explorar e comparar como técnicas de diferentes naturezas lidam com o problema da esparsidade e de alta dimensionalidade no problema de coagrupamento, explicando quais dificuldades encontram e como as superam. Assim, a questão de pesquisa que delimita esta pesquisa está relacionada às dificuldades enfrentadas por algoritmos de coagrupamento quando eles são aplicados a dados textuais, que possuem uma representação esparsa, e como tais dificuldades são superadas pelos algoritmos que obtêm os melhores desempenhos.

A fim de discutir a pesquisa delineada em coagrupamento de textos, este artigo segue organizado da seguinte forma: a Seção 2 apresenta o problema de coagrupamento aplicado a dados textuais; a Seção 3 descreve a abordagem proposta nesta pesquisa para tratar o problema; a Seção 4 detalha o processo de avaliação dos resultados; a Seção 5 descreve as atividades já realizadas, e a Seção 6 apresenta a conclusão.

2. APRESENTAÇÃO DO PROBLEMA

Em mineração de textos, é importante trabalhar para melhoria dos resultados de análise subjetivas que trazem significado interessante para o contexto do qual os dados derivam. Uma das possibilidades para alcançar essa melhoria é fazer uma análise de similaridade parcial, ou seja, não usar todos os atributos descritivos do dado no cálculo da similaridade (o que não é feito na análise de agrupamento). Essa é uma análise capaz de fornecer resultados diferenciados e potencialmente úteis para alguns contextos ou objetivos específicos.

A análise de similaridade parcial é uma das características inerentes aos processos de coagrupamento porque encontra subconjuntos de objetos e atributos da base de dados que reflitam relações significativas entre si. Dessa forma, a análise dos dados é feita apenas em sub-regiões da base original, realizando de forma intrínseca o procedimento de eliminar atributos redundantes, ou de pouca relevância, de acordo com o conjunto de objetos que está sendo avaliado no momento [6].

Mas a tarefa de coagrupamento envolve um problema computacional complexo, segundo [6], na forma mais simples, quando A é uma matriz binária, um *bicluster* corresponde a um *biclique* de um grafo bipartido, supondo que as linhas da matriz A são os vértices à esquerda e as colunas os vértices à direita, e os valores dos elementos indicam se existe uma aresta ligando uma linha a uma coluna, caso o valor seja 1, ou 0, caso contrário. Encontrar o *bicluster* ou cogrupo de tamanho máximo, nessa situação, é equivalente a encontrar o *biclique* com o máximo de arestas em um grafo bipartido, que é um problema NP-completo [14]. Assim, é possível inferir que problemas mais complexos nos quais a matriz A apresenta valores numéricos reais que serão levados em conta para calcular a qualidade de um *bicluster*, não apresentará

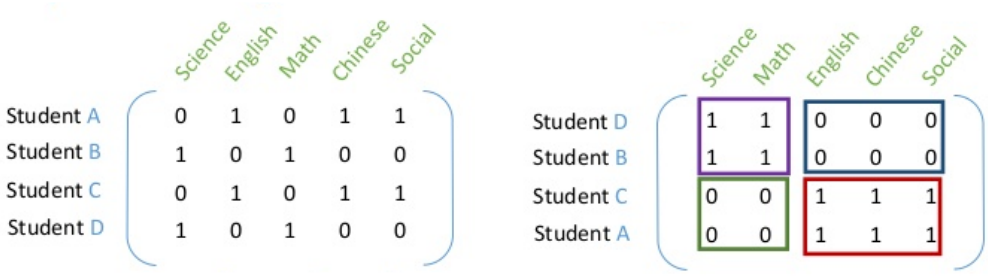


Figure 1: Exemplo de coagrupamento de um conjunto de dados

complexidade menor do que a forma mais simples [11, 3].

Por este motivo, grande parte dos algoritmos de coagrupamento são métodos inexatos, baseados em heurísticas ou meta-heurísticas. O problema tratado nesta pesquisa é a exploração, análise e comparação de técnicas heurísticas, capazes de fornecer soluções sub-ótimas para o problema em tempo aceitável, que sejam adequadas para a realização da busca de soluções em espaços caracterizados pela esparsidade e a alta dimensionalidade dos dados. O objetivo é construir um estudo que explique os aspectos positivos e negativos do uso de diferentes algoritmos nesse ambiente de experimentação.

3. PROPOSTA DE SOLUÇÃO

Esta pesquisa exige um estudo exploratório sobre os algoritmos de coagrupamento apresentados na literatura assim como sobre as principais métricas utilizadas para avaliação dos resultados produzidos no processo de busca pelos cogrupos. Assim, os algoritmos poderão ser aplicados a dados textuais em um ambiente de experimentação sistemático que permita construir avaliações detalhadas tanto em termos quantitativos quanto qualitativos.

Uma vez que o contexto de análise diz respeito a dados textuais, antes de aplicar os algoritmos, será necessário proceder com uma série de rotinas de pré-processamento. Serão utilizadas rotinas de pré-processamento comuns na área de mineração de texto tais como: *tokenização*, filtragem de *stopwords*, e representação da relação "termos × documento" usando representações binárias e *tf–idf* (*term frequency – inverse document frequency*) [16].

Os algoritmos que foram escolhidos até o momento abrangem três diferentes formas de resolver o problema de coagrupamento. São eles:

- Algoritmo de Cheng & Church [4]: Este algoritmo encontra *biclusters* que sejam tão grandes quanto possível (máximo volume), minimizando o Resíduo Quadrático Médio (medida de homogeneidade do *bicluster*) destes. O algoritmo usa uma estratégia gulosa.
- Algoritmo bicACO [7]: Este algoritmo apresenta uma meta heurística com a finalidade de melhorar a estratégia gulosa, ao eliminar linhas e colunas do algoritmo de *Cheng & Church*. O algoritmo bicACO é a primeira técnica de biagrupamento a usar os princípios da otimização de colônias de formigas (Ant Colony Optimization) para resolver o problema do biagrupamento.
- Algoritmos baseados em fatoração de matrizes não-negativas (NMF) [10]: O problema de fatoração de matrizes não-negativas é considerado um método de agrupamento e tem sido desenvolvido como um método de redução de dimensionalidade e estendido para o problema de coagrupamento. Ele segue processos de otimização que podem ser implementados por diferentes métodos.

No decorrer da pesquisa outros algoritmos poderão ser incorporados ao estudo. Em um primeiro momento, as experimentações serão executadas em um ambiente de teste controlado, representado por uma coleção de conjunto de dados sintéticos. Na sequência, o ambiente de experimentação receberá os dados textuais, em língua portuguesa.

4. AVALIAÇÃO DA SOLUÇÃO

A avaliação da qualidade dos cogrupos detectados é um problema importante e desafiador, independente do tipo de dado que está sendo utilizado no experimento, devido aos diferentes critérios utilizados e aos diferentes objetivos relacionados ao coagrupamento. Há avaliações estabelecidas para análise dos cogrupos formados, porém, há também avaliações que dizem respeito ao resultado de agrupamento que pode ser derivado do processo de coagrupamento[1].

O resultados dos algoritmos serão avaliados sob quatro aspectos: avaliação interna, que avalia a estrutura intrínseca dos dados expressa pelos cogrupos (índices Silhouette e Dunn); avaliação externa, que mede a precisão com que os cogrupos (de linhas) correspondem à estruturas que sabe-se existir nos dados (índices Rand e Informação Mútua); inspeção visual sobre representações gráficas obtidas em problemas com matrizes esparsas sintéticas; avaliação do conteúdo semântico que pode ser extraído dos cogrupos de textos [9].

5. ATIVIDADES JÁ REALIZADAS

Até o presente momento, foram realizados estudos exploratórios sobre algoritmos de agrupamento e coagrupamento. Tais estudos foram feitos para construir um panorama sobre as pesquisas mais recentes na área, de forma a conhecer os objetivos, estratégias e avaliações usadas pelos pesquisadores. Como resultado dessa atividade, os trabalhos correlatos à essa pesquisa foram organizados e os algoritmos a serem testados foram escolhidos. Também, todas as rotinas de pré-processamento de textos já foram estudadas e se encontram

[1]O processo de coagrupamento gera os chamados "grupos de linhas" que, na realidade, podem ser vistos como os clássicos grupos obtidos nos processos de agrupamento. Nesse caso, o coagrupamento é visto como um processo diferenciados para resolver a tarefa de agrupamento.

prontas para aplicação. Atualmente, os algoritmos a serem analisados estão em fase de implementação.

6. CONCLUSÃO

Este projeto tem como objetivo principal apresentar uma comparação de estratégias heurísticas aplicadas ao problema de coagrupamento de dados textuais, fazendo uma análise quantitativa e qualitativa dos resultados produzidos, em termos de métricas específicas de avaliação de coagrupamento, e de avaliação de produção da informação semântica útil para análise textual.

Enfim, espera-se responder à questão de pesquisa, contribuir com a área de mineração de dados em termos de: a) aprofundamento dos conceitos de coagrupamento aplicado a dados textuais; b) entendimento claro da influência que a esparsidade e a alta dimensionalidade exerce sobre cada um desses algoritmos que foram escolhidos para serem estudados, de forma que seja possível propor uma alteração em características ou aspectos específicos ao um algoritmo para alcançar a melhoria neste tipo de dado; c) apresentar algumas heurísticas de coagrupamento para o análise de dados textuais; d) apresentar avaliação das heurísticas de coagrupamento aplicadas em dados textuais; d) estabelecer comparativos que mostram que o uso de coagrupamento traz benefícios em termos de qualidade na análise de texto.

7. REFERÊNCIAS

[1] A. Amine, Z. Elberrichi, and M. Simonet. Evaluation of text clustering methods using wordnet. *Int. Arab J. Inf. Technol.*, 7(4):349–357, 2010.

[2] P. Berkhin. *A Survey of Clustering Data Mining Techniques*, pages 25–71. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[3] L. Bulteau, V. Froese, S. Hartung, and R. Niedermeier. Co-clustering under the maximum norm. *Algorithms*, 9(1):17, 2016.

[4] Y. Cheng and G. M. Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.

[5] L. A. da Silva, S. Peres, and C. Boscarioli. *Introdução À Mineração de Dados - Com Aplicação Em R.* Elsevier Editora Ltda., São Paulo, Brasil, 2016.

[6] F. O. de Franca. *Biclusterização na Análise de Dados Incertos.* PhD thesis, Universidade Estadual de Campinas, 2010.

[7] F. O. de França, G. P. Coelho, and F. J. Von Zuben. bicaco: An ant colony inspired biclustering algorithm. In *International Conference on Ant Colony Optimization and Swarm Intelligence*, pages 401–402. Springer, 2008.

[8] R. Feldman and J. Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge university press, 2007.

[9] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[10] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[11] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1):24–45, 2004.

[12] R. M. Marcacini and S. O. Rezende. Incremental construction of topic hierarchies using hierarchical term clustering. In *SEKE*, page 553, 2010.

[13] G. Mi, Y. Gao, and Y. Tan. Apply stacked auto-encoder to spam detection. In *International Conference in Swarm Intelligence*, pages 3–15. Springer, 2015.

[14] R. Peeters. The maximum edge biclique problem is np-complete. *Discrete Applied Mathematics*, 131(3):651 – 654, 2003.

[15] R. G. Pensa, D. Ienco, and R. Meo. Hierarchical co-clustering: off-line and incremental approaches. *Data Mining and Knowl. Disc.*, pages 1–34, 2014.

[16] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975.

[17] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

Uma abordagem dinâmica para avaliação da serendipidade de Sistemas de Recomendação.

Alternative Title: A dynamic approach for evaluating serendipity of Recommender Systems.

André Paulino de Lima
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
andre.p.lima@usp.br

Sarajane Marques Peres
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
sarajane@usp.br

RESUMO

A literatura recente de Sistemas de Recomendação (SR) sobre métodos de avaliação *off-line* propõe estimar a serendipidade de um SR por meio de uma medida de surpresa. Nestes métodos, o SR avaliado é mantido estático durante a aplicação da medida. Entretanto, a medida de surpresa de um item varia conforme o usuário é exposto a novas recomendações. Um método que não considera a variação da surpresa ao longo do tempo possui, em princípio, uma capacidade reduzida de usar a informação disponível no SR para estimar a serendipidade de futuras recomendações.

Este projeto propõe uma medida e um método de avaliação *off-line* que simula sequências de interações de usuários com um Sistema de Recomendação Baseado em Conteúdo (SRbC), no intuito de descrever a distribuição esperada da surpresa de futuras recomendações. Conhecer tais perfis permitiria personalizar a calibração entre objetivos de um SR, como relevância e surpresa, de maneira mais precisa. Resultados obtidos com a aplicação do método proposto sobre diferentes configurações de um SRbC são comparados com resultados obtidos pela aplicação de um método de avaliação *off-line* no estado da arte sobre as mesmas configurações, com o intuito de identificar vantagens e desvantagens entre as duas abordagens.

Palavras-Chave

Sistemas de Recomendação, avaliação, serendipidade, surpresa

ABSTRACT

Recent works in Recommender Systems (RS) literature on off-line evaluation methods propose estimating the serendipity of an SR by means of a measure of surprise. In such methods, the experimenter keeps the RS static while the measure is applied. However, the surprise attributed to an item varies as the RS presents new recommendations to its users. In principle, a method that does not consider surprise variation over time has a reduced capacity to employ information available in the RS to estimate serendipity of future recommendations.

This project proposes a measure and an off-line evaluation method that simulates sequences of user interactions with a Content-Based Recommender System (CBRS) in order to describe the expected distribution of surprise of future recommendations. Knowing such profiles would allow personalizing the balance between objectives of a SR, as relevance and surprise, in a more precise way. Results obtained by applying the proposed method on different configurations of a CBRS are compared to those obtained by applying a state-of-the-art off-line evaluation method on the same sonfigurations, in order to identify advantages and disadvantages between the two approaches.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•**Information systems** → **Recommender systems;** *Novelty in information retrieval; Information retrieval diversity;*

Keywords

Recommender Systems; evaluation; serendipity, surprise

1. INTRODUÇÃO

Nos primeiros trabalhos da área de pesquisa em Sistemas de Recomendação (SRs), o desempenho desses sistemas era medido predominantemente pela sua **acurácia** em antecipar quais itens seriam do interesse de seus usuários. Atualmente, apesar da acurácia continuar sendo uma propriedade crítica para SRs, a literatura identifica outras propriedades desejáveis, como cobertura, diversidade e serendipidade.

Foco deste trabalho, a **serendipidade** é uma propriedade difícil de ser definida devido à subjetividade que envolve o conceito. Uma das definições mais antigas para serendipidade disponível na literatura tem caráter informal: uma recomendação serendipitosa apresenta um item surpreendentemente interessante e que talvez nunca fosse encontrado por uma busca conduzida pelo usuário [6]. Entretanto, a literatura recente assume que serendipidade possa ser decomposta em diferentes fatores, como **surpresa** e relevância, para os quais existem definições e medidas objetivas [2, 6].

Os métodos de avaliação *off-line* no estado da arte estimam a serendipidade de um SR por meio da surpresa observada em uma amostra de recomendações produzidas pelo sistema, o qual é mantido estático durante a avaliação. Contudo, tais métodos não capturam a tendência de variação da surpresa por não considerarem a variação da surpresa associada a um item conforme o SR expõe seus usuários a novos itens. Conhecer a tendência de variação da surpresa permitiria comparar como diferentes configurações de um SR distribuem recomendações serendipitosas ao longo do tempo.

2. APRESENTAÇÃO DO PROBLEMA

Em uma revisão recente das medidas para surpresa [8], é possível verificar que estas mantêm algumas semelhanças importantes entre si:

1. Compartilham a intuição de que a surpresa de um item novo é inversamente proporcional ao quanto este é similar aos itens que já foram recomendados ao usuário.

2. A modelagem da medida de surpresa de um item envolve a definição de uma função que mede a similaridade entre este item e um item ao qual o usuário foi exposto.

3. A modelagem da medida de surpresa envolve alguma forma de redução[1] dos diferentes valores de similaridade obtidos entre o item novo e cada um dos elementos do conjunto de itens já observados pelo usuário.

Decorre da intuição descrita no primeiro item que a surpresa não é uma propriedade inerente ao item, e sim uma propriedade que emerge da exposição do usuário aos itens no repositório do SR. Por esta razão, um mesmo item pode ser associado a diferentes graus de surpresa por usuários diferentes. Na verdade, mesmo do ponto de vista de um único usuário, a surpresa de um item é uma propriedade variável: um mesmo item pode ser associado a diferentes graus de surpresa conforme o usuário é exposto a novas recomendações.

Métodos de avaliação *off-line* no estado da arte, como *one plus random* [3, 7, 8], estimam a serendipidade de um SR por meio da surpresa observada em uma amostra de recomendações produzidas pelo sistema, que é mantido em um estado fixo durante a avaliação. Entretanto, como já comentado, a surpresa tem um caráter variável, observado durante a dinâmica da interação entre o sistema e seus usuários. Logo, para estimar a tendência de um SR em gerar recomendações serendipitosas, é necessário considerar essa variabilidade.

3. PROPOSTA DE SOLUÇÃO

O método proposto consiste na combinação de duas capacidades: (a) uma medida que compute a surpresa máxima que o SRbC pode oferecer a um usuário, levando em consideração sua exposição aos itens do repositório do SR (seção 3.1), e (b) um método que avalie a surpresa de sequências de estados de um SRbC (seção 3.2). Entretanto, antes de elaborar as capacidades mencionadas, é necessário descrever os componentes típicos da arquitetura de SRbCs [4] adotados neste desenvolvimento conceitual:

[1]Redução no sentido de combinar, agregar ou sumarizar diferentes valores em um único valor, como no sentido empregado no paradigma *map-reduce*.

1. Analisador de Conteúdo: $g_c : I \Rightarrow V$, sendo I o conjunto de itens no repositório do SR e V o conjunto com a representação numérica desses itens (por exemplo, representação vetorial, com $V \subset \mathbb{R}^m$). Para fins de simplificação, assumimos que I é um conjunto finito.

2. Construtor de Perfil: $g_p : T_r \Rightarrow \Theta$, sendo T_r o conjunto de treinamento, composto por itens que possuem avaliação, e sendo Θ o conjunto de modelos induzidos para representar os perfis de usuários. O conjunto de treinamento T_r é um subconjunto de $V \times U \times R$, sendo U o conjunto de usuários e R o domínio de valores de avaliação (*rating*) que um item em V pode receber.

3. Filtro de Conteúdo: $g_f : V \times U \times \Theta \Rightarrow R$, é uma função que estima a avaliação que um usuário $u \in U$ atribuiria a um item $v \in V$ em um determinado momento, uma vez que Θ_u pode variar conforme o usuário é exposto a novos itens ao longo do tempo.

3.1 A medida de surpresa potencial

O objetivo da medida aqui proposta é determinar um limite superior da surpresa que um SR pode oferecer a um usuário, considerando os itens aos quais este foi exposto. A ideia é explorar a ordem na qual o sistema recomenda os itens ainda não observados pelo usuário, pois esta influi na surpresa total que o usuário pode perceber.

Seja E_u o subconjunto de I que representa os itens aos quais o usuário u foi exposto e N_u o subconjunto de itens dos quais o usuário u não tem conhecimento ($N_u = I - E_u$). A **surpresa potencial** de N_u dado E_u é:

$$S_p(N_u, E_u) = \max_{seq \in permut(N_u)} S_s(seq, E_u)$$

sendo seq uma permutação dos itens em N_u, representando uma sequência possível de apresentação dos itens em N_u ao usuário u, e $S_s(seq, E_u)$ corresponde à soma da surpresa observada em cada item contido na sequência seq. Em outras palavras, o predicado $S_s(seq, E_u)$ computa a surpresa do primeiro item da sequência em relação aos itens em E_u; em seguida, o primeiro item é incorporado ao conjunto E_u e computa-se a surpresa do segundo item em seq, e assim sucessivamente até que o último elemento de seq seja processado. Desta forma, a **surpresa de uma sequência** seq dado Eu é definida como:

$$S_s(seq, E_u) = \begin{cases} 0, & \text{se } |seq| = 0, \\ S_i(h, E_u) + S_s(t, E_u \cup \{h\}), & \text{se } |seq| > 0. \end{cases}$$

sendo h ($head$) o primeiro elemento de seq e t ($tail$) o restante dos elementos em seq. Esta definição faz uso de um outro predicado, $S_i(h, E_u)$, que computa a surpresa de um único item em relação ao conjunto E_u. Assim, a **surpresa de um item** dado E_u é definida como:

$$S_i(item, E_u) = \min_{item' \in E_u} s(item, item')$$

sendo $s(item, item')$ uma medida de surpresa do $item$ em relação ao $item'$. Por exemplo, se os itens estiverem representados como vetores ($V \subset \mathbb{R}^m$), a medida de surpresa

$s(i, j)$ poderia ser implementada como a distância euclidiana. Outras medidas de surpresa, como aquelas listadas no Quadro 1, também poderiam ser utilizadas como $s(i, j)$.

Da definição de surpresa potencial $S_p(N_u, E_u)$, decorre que esta é uma quantidade finita e é possível interpretá-la como sendo a maior surpresa possível que o SR pode oferecer ao usuário u, dada sua exposição E_u; ou de forma mais intuitiva: é uma medida do estoque de surpresa armazenada em um estado do sistema, representado por N_u.

Entretanto, encontrar a sequência de recomendações que maximiza a surpresa potencial, como define o predicado $S_p(N_u, E_u)$, é um problema de complexidade exponencial. A busca exaustiva não é viável quando a escala do problema cresce e, nesta situação, a prática sugere a adoção de métodos aproximados compatíveis com a escala do problema.

3.2 A simulação da dinâmica de um SRbC

A medida de surpresa potencial proposta se aplica a um estado do SR. Contudo, como discutido em [1], a avaliação de propriedades dinâmicas de SRs demanda uma abordagem que considere sua variabilidade. Por esta razão, o objetivo do algoritmo descrito a seguir é simular a interação do usuário com o sistema, de modo a permitir observar a variação da surpresa potencial. Esta simulação consiste basicamente em três passos: (a) computa-se a surpresa potencial do sistema, observando a exposição atual do usuário; (b) usando o perfil do usuário (g_f), o sistema seleciona um item e o apresenta ao usuário; este item é, portanto, incorporado ao conjunto de itens aos quais o usuário foi exposto, e (c) os passos anteriores se repetem até que não haja novos itens a recomendar. Ao final da simulação, uma curva com a evolução da surpresa potencial pode ser desenhada.

Assumindo que as interações de um usuário com o SR não afetem as interações dos demais usuários (o que é verdade para SRbC), é possível descrever os passos acima nos termos empregados na seção anterior, como segue:

1. Seja $r : U \times I \Rightarrow R$ uma função que mapeia a avaliação que um usuário $u \in U$ atribui ao item $i \in I$. Assume-se que todos os itens em E_u foram avaliados pelo usuário. Para maior clareza da notação, a expressão r_{ui} é usada no lugar de $r(u, i)$.

2. Seja Θ_u um modelo induzido a partir de E_u, representando a preferência do usuário u. Assume-se que Θ_u aproxime a distribuição probabilística condicional da preferência do usuário, nominalmente $p(r_{ui}|i)$.

3. Seja L o item em N_u que obtém a maior pontuação segundo o filtro de conteúdo (g_f). A simulação se inicia com o SRbC expondo o item L ao usuário u.

4. Com probabilidade p, o usuário u seleciona o item L e lhe atribui como avaliação a pontuação estimada por g_f. Com probabilidade $(1 - p)$, o usuário u seleciona um item em N_u diferente de L e atribui-lhe uma avaliação conforme uma distribuição probabilística Q.

5. Uma aproximação da surpresa potencial $S_p(N_u, E_u)$ é computada e salva, e o item selecionado pelo usuário u no passo anterior é removido de N_u e incorporado a E_u.

6. Os passos de 3 a 6 se repetem até que o estoque de surpresa potencial tenha se reduzido a um nível crítico (S_{lb}) ou k itens tenham sido avaliados pelo usuário u. Além disso, a cada l_θ iterações, o modelo Θ_u é atualizado.

7. A tendência de variação de surpresa do SR é a diferença entre a área sob a curva de surpresa potencial e a área do menor retângulo no qual aquela curva se inscreve.

A proposta apresentada se vale de diversos parâmetros para descrever a simulação: alguns pertencem ao próprio SRbC (como Θ_u), enquanto outros representam características do ambiente no qual se dá a simulação (como p, Q, S_{lb}, k e l_θ). Estes parâmetros serão estimados a partir do conjunto de dados adotado para avaliação da solução.

Como mencionado anteriormente, métodos no estado-da-arte, como o *one plus random* estimam a surpresa de um SR por meio da avaliação da média da surpresa obtida em uma amostra de recomendações produzidas pelo sistema (em [7], a amostra é composta por 10 itens por usuário de teste), contra uma versão estática do perfil do usuário (E_u não varia durante a avaliação). Por esta razão, a estimativa produzida reflete apenas o potencial de surpresa do sistema em seu estado atual. A contribuição do método proposto consiste em produzir uma estimativa que não reflita somente o estado atual, mas também o perfil de variação da surpresa esperado. Este perfil permitiria, por exemplo, analisar a tendência do sistema em concentrar recomendações serendipitosas no curto prazo ou em distribuí-las de forma mais esparsa. Ou ainda, obter uma curva que represente o limite superior ou inferior para a surpresa (por meio de modificação da política de construção do perfil, otimizado para surpresa ou relevância, respectivamente). De um ponto de vista prático, o método proposto oferece estimativas mais precisas que podem ser usadas para personalizar a calibração de objetivos de um SR, como priorizar relevância ou surpresa.

4. PROJETO DE AVALIAÇÃO

O conjunto de dados MovieLens-1M [5] será utilizado para avaliar o método proposto. Este conjunto de dados foi escolhido por apresentar um balanço interessante entre avaliações (mais de 1 milhão), usuários (6.040) e itens (3.706). Este último é relevante por conta da natureza exponencial da medida de surpresa potencial proposta. Com o intuito de antecipar possíveis problemas no projeto, o trabalho inclui em 3 experimentos iniciais:

1. Ambiente controlado: o experimento consiste em implementar e validar um ambiente no qual o comportamento dinâmico de um SRbC possa ser simulado, conforme descreve a seção 3.2. Um conjunto de dados fictício foi construído para permitir a validação do ambiente antes de sua aplicação aos dados do MovieLens-1M.

2. Representações de Texto: o objetivo é identificar parâmetros do processo de representação vetorial que produzem maior impacto no desempenho de uma tarefa de avaliação de similaridade semântica, como *SemEval-2017 Task 1*. Serão avaliadas representações vetoriais tradicionais (*tf-idf*) e representação distribuída de texto [9].

3. Simulação: o experimento consiste em avaliar o impacto de diferentes representações de texto na tendência do SRbC em expressar serendipidade. Durante a avaliação, curvas de surpresa potencial serão computadas e comparadas para diferentes configurações. Sob as mesmas configurações, será aplicado como base de comparação um método *off-line* no estado da arte (p. ex.: *one plus random*) para identificar possíveis situações nas quais os métodos alcançam conclusões distintas.

Quadro 1: Medidas de surpresa propostas na literatura.

Medida	Descrição	Intuição
Surpresa como uma função de distância [7]	$s(i,j) = 1 - \frac{\|L_i \cap L_j\|}{\|L_i \cup L_j\|}$ onde: L_i é o conjunto de *labels* associado ao item i.	A surpresa atribuída a um item i em relação a um item j (já conhecido pelo usuário) é proporcional à distância entre as representações numéricas dos itens i e j. Nesta medida, os autores assumem que existe um conjunto de *labels* associado a cada item (por exemplo, gêneros de um filme). A distância de Jaccard é usada para computar a surpresa entre os itens.
Surpresa como uma função da distribuição probabilística (ocorrência conjunta dos itens) [7]	$s(i,j) = -log_2 \frac{p(i,j)}{p(i)p(j)} / log_2 p(i,j)$ onde: $p(i)$ é a probabilidade do item i ter sido avaliado por qualquer usuário. $p(i,j)$ é a probabilidade dos itens i e j terem sido avaliados por um mesmo usuário.	A surpresa atribuída a um item i em relação a um item j (já conhecido pelo usuário) depende da frequência com a qual os itens são avaliados individual e conjuntamente. Os autores assumem que existe um conjunto de usuários associado a cada item (usuários que avaliaram o item). A medida de associação *normalized pointwise mutual information* é usada para computar a surpresa entre os itens.

5. ATIVIDADES JÁ REALIZADAS

Uma versão inicial do ambiente controlado para simulação do comportamento dinâmico de um SRbC foi construída, junto com um conjunto de dados fictício para validá-lo. Curvas de surpresa potencial foram obtidas para este conjunto de dados fictício, usando uma aproximação gulosa para a medida de surpresa potencial. Além disso, curvas de surpresa potencial foram computadas para três usuários no conjunto MovieLens-1M: 1) para o usuário que possui o maior número de avaliações (2.303); 2) para um dos usuários com o menor número de avaliações (19) e 3) para um usuário com número médio de avaliações (168).

Com relação à preparação do conjunto MovieLens-1M, este foi enriquecido com descrições de filmes obtidas no site MovieLens (movielens.org), mantido pelo mesmo time de pesquisa que publicou o conjunto de dados.

As atividades em execução consistem em: (a) avaliar possíveis alternativas para aproximação da medida de Supresa Potencial (como programação dinâmica, A* ou algoritmos genéticos); (b) converter os itens no conjunto MovieLens-1M em representações vetoriais considerando o texto associado a cada item e (c) estimar os parâmetros que serão utilizados pelo ambiente de simulação em sua versão final.

6. CONCLUSÃO

Neste trabalho está sendo proposto um método de avaliação *off-line* da serendipidade que explora o aspecto variável da surpresa. O método simula o comportamento dinâmico de um SRbC para capturar a variação da surpresa potencial do sistema enquanto os usuários de teste são expostos a novas recomendações. A contribuição principal do trabalho é o desenvolvimento de um método capaz de estimar a tendência de um SRbC em distribuir recomendações serendipitosas: concentradas no curto prazo ou dispersas ao longo do tempo. O método proposto pode empregar medidas de surpresa recentemente propostas na literatura. Ainda, o método proposto é explorado para avaliar como o uso de técnicas emergentes de representações vetorial de texto impactam na tendência do sistema em distribuir recomendações serendipitosas no tempo.

7. REFERÊNCIAS

[1] R. Burke. Evaluating the dynamic properties of recommendation algorithms. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 225–228. ACM, 2010.

[2] P. Castells, N. J. Hurley, and S. Vargas. Novelty and diversity in recommender systems. In F. Ricci, L. Rokach, and B. Shapira, editors, *Handbook of Recommender Systems*, pages 881–905. Springer, New York, NY, 2015.

[3] P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46. ACM, 2010.

[4] M. de Gemmis, P. Lops, C. Musto, F. Narducci, and G. Semeraro. Semantics-aware content-based recommender systems. In F. Ricci, L. Rokach, and B. Shapira, editors, *Handbook of Recommender Systems*, pages 119–159. Springer, New York, NY, 2015.

[5] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2015.

[6] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[7] M. Kaminskas and D. Bridge. Measuring surprise in recommender systems. In *Proceedings of the Workshop on Recommender Systems Evaluation: Dimensions and Design (Workshop Programme of the 8th ACM Conference on Recommender Systems)*. Citeseer, 2014.

[8] M. Kaminskas and D. Bridge. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):2, 2016.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

MOOC como um software colaborativo: um estudo exploratório dos requisitos para suporte à abordagem conectivista sob a ótica do Modelo 3C

Alternative Title: MOOC as collaborative software: an exploratory study of requirements to support the connectivist approach from the perspective of 3C Model

Neyla Teixeira Fontan
Universidade Federal da Bahia
Salvador, Bahia
neylafontan@gmail.com

Rita Suzana Pitangueira Maciel
Universidade Federal da Bahia
Salvador, Bahia
ritasuzana@dcc.ufba.br

RESUMO

Massive Open Online Course (MOOC) tem sido apontado como uma tendência na educação à distância, através da oferta de cursos via web (*online*) com acesso aberto (*open*) para quantidade ilimitada de participantes (*massive*). Diferentes classificações são atribuídas a projetos de MOOC, mas a literatura converge em dividi-los em dois grupos: cMOOC e xMOOC. O tipo cMOOC baseia-se no Conectivismo, considerado por seus autores como a teoria de aprendizagem para a era digital. No tipo xMOOC, a abordagem pedagógica está mais próxima à utilizada em cursos online tradicionais. De acordo com a Teoria Conectivista, a aprendizagem emerge a partir de uma rede de conexões entre aprendizes, monitores e tutores que interagem de forma colaborativa. Pensando em MOOC como um sistema colaborativo no contexto da engenharia de software, observa-se que o conectivismo é pouco explorado e não há um consenso sobre os requisitos que guiem a construção de softwares do domínio cMOOC. Esta dissertação de mestrado propõe a realização de um estudo exploratório com objetivo de identificar os requisitos, sob a ótica do Modelo 3C de Colaboração, que atendam à abordagem conectivista. Com o resultado do estudo exploratório, espera-se criar um *framework* conceitual que apoie o processo de engenharia de requisitos para implementação de softwares do domínio cMOOC.

Palavras-Chave

MOOC, Conectivismo, Sistemas Colaborativos, Engenharia de Requisitos, Modelo 3C de Colaboração.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

ABSTRACT

Massive Open Online Course (MOOC) has been pointed out as a trend in distance education, through the offering of web courses (online) with open access (open) for unlimited number of participants (massive). Different classifications are assigned to MOOC projects, but the literature converges to divide them into two groups: cMOOC and xMOOC. The cMOOC type is based on Connectivism, considered by its authors as the learning theory for the digital age. In the xMOOC type, the pedagogical approach is closer to that used in traditional online courses. According to the Connectivist Theory, learning emerges from a network of connections between learners, monitors and tutors who interact in a collaborative way. Thinking about MOOC as a collaborative system in the context of software engineering, it is observed that the connectivism is little explored and there is no consensus on the requirements that guide the development of cMOOC domain software. This dissertation proposes the accomplishment of an exploratory study aiming to identify the requirements, from the perspective of 3C Collaboration Model, that attend to the connectivist approach. With the result of the exploratory study, it is expected to create a conceptual framework that supports the requirements engineering process for the implementation of cMOOC domain software.

CCS Concepts

•Information systems → Collaborative and social computing systems and tools; •Software and its engineering → Requirements analysis; •Applied computing → *Education;* **•Human-centered computing** → *Computer supported cooperative work;*

Keywords

MOOC, Connectivism, Collaborative Systems, Requirements Engineering, 3C Collaboration Model.

1. INTRODUÇÃO

O termo *Massive Open Online Course* (MOOC) foi cunhado em 2008 para descrever um modelo de curso online

desenvolvido pelos acadêmicos Stephen Downes e George Siemens na Universidade de Manitoba, no Canadá [1]. Esse curso foi projetado com base em princípios conectivistas, nos quais a aprendizagem e o conhecimento emergem de uma rede de conexões [7]. O curso congregou 2.300 (dois mil e trezentos) participantes online que o acessaram gratuitamente. Esse curso e seus sucessores da linha conectivista deram origem ao modelo cMOOC ou *Connectivist* MOOC [7].

Depois desse início, houve uma rápida expansão com o surgimento diversos de provedores e plataformas tecnológicas para oferta de MOOC, contudo nem todos seguiram a abordagem conectivista [1] proposta originalmente. O ano de 2012 foi marcado pelo lançamento de provedores MOOC vinculados à renomadas instituições, tais como o provedor *Coursera* da Universidade Stanford e a plataforma *edX* lançada em parceria entre as universidades MIT e Harvard. Naquele ano, surgiram também o *Udacity* e a britânica *FutureLearn* [7]. Esses provedores de MOOC deram origem ao modelo extensionista ou xMOOCs, com abordagem pedagógica mais próxima à utilizada por cursos online tradicionais. O conteúdo do curso é previamente definido e permanece inalterado ao longo de sua execução. São cursos baseados em vídeos com aulas expositivas e com avaliação através de questões de múltipla escolha [10].

Diferentes classificações surgiram posteriormente e são atribuídas a projetos de MOOC, mas a literatura converge em dividi-los entre cMOOC e xMOOC [4].

A Teoria Conectivista que embasa o projeto de criação de um cMOOC é considerada por seus autores, Downes e Siemens, como a teoria de aprendizagem para a era digital [7]. Siemens entende que a tecnologia reorganizou o nosso modo de viver, de comunicar e de aprender. Para o autor, a aprendizagem ocorre de várias formas, incluindo a aprendizagem informal alcançada através de comunidades de prática e de redes sociais [11].

Um dos princípios que caracterizam o conectivismo é o incentivo à participação de aprendizes em atividades realizadas colaborativamente com outros colegas aprendizes, monitores e tutores [1]. A construção de conteúdo do curso e a avaliação de exercícios, entre outras atividades, são estimuladas a ocorrer de forma colaborativa entre os participantes.

A colaboração está no cerne da concepção de um projeto de MOOC conectivista, sendo o seu principal requisito para desenvolvimento de softwares que atendam ao domínio cMOOC.

Pensando em MOOC como um sistema colaborativo no contexto da engenharia de software, observa-se que não há um conjunto de requisitos que guiem os engenheiros de software na construção de plataformas tecnológicas para disponibilização de cursos [4]. A abordagem conectivista é pouco explorada e não há um consenso sobre os requisitos que deem suporte às práticas do conectivismo para desenvolvimento de softwares do domínio cMOOC [2].

2. APRESENTAÇÃO DO PROBLEMA

Enquanto os ambientes de EAD tradicionais são construídos a partir de customização dos requisitos de consenso para esse domínio, não há uma definição sobre os requisitos que um software do domínio cMOOC deva atender. Assim, projetos de construção de cMOOC são sempre iniciados do zero, passando por todas as fases do processo de Engenharia de Requisitos, o que demanda tempo e custo ao projeto.

O problema apresentado leva à seguinte questão de pesquisa que guia este estudo: quais requisitos devem ser atendidos por um software do domínio MOOC para suporte à abordagem conectivista?

3. PROPOSTA DE SOLUÇÃO

Esta pesquisa propõe realizar um estudo exploratório para identificar um conjunto de requisitos que guiem a construção de softwares para o domínio MOOC conectivista.

Considerando que a prática do conectivismo é fortemente apoiada por atividades realizadas colaborativamente entre os participantes, um MOOC conectivista pode ser analisado como um sistema colaborativo no contexto da engenharia de software. Segundo Ellis et al. [3], sistemas colaborativos, também identificados na literatura como *Groupware* ou *Computer Supported Cooperative Work* (CSCW), são sistemas computacionais utilizados para apoiar o trabalho em grupo com objetivo de atingir um determinado fim.

A partir de um refinamento dos conceitos sobre sistemas colaborativos apresentados no modelo de Ellis et al. [3], o grupo de pesquisa de Pimentel et al. [9], propôs o Modelo 3C de Colaboração para análise de sistemas colaborativos, com algumas diferenças de nomenclatura. O Modelo 3C de Colaboração analisa um sistema colaborativo sob três dimensões: comunicação, coordenação, cooperação; e pela percepção do que ocorre em cada dimensão [5].

A dimensão *comunicação* refere-se à troca de mensagens entre os participantes; a *coordenação* consiste no gerenciamento de pessoas, tarefas e recursos; e a *cooperação* representa a atuação conjunta dos participantes para produção de objetos ou informações em um espaço compartilhado [9]. Através da *percepção* do que ocorre em cada uma das três dimensões, os participantes acompanham suas ações e as ações de seus colegas [5]. Os autores entendem que a colaboração é caracterizada pela realização de todo o trabalho em conjunto, o que envolve comunicação, coordenação, cooperação e a percepção sobre o que acontece nessas dimensões [6].

Pimentel et al. [9] acreditam que o Modelo 3C de Colaboração tem sido apropriado para o desenvolvimento de sistemas colaborativos [9]. Eles avaliam que o modelo é útil em diferentes etapas do processo de desenvolvimento, como por exemplo, na análise do *groupware* a ser desenvolvido e no desenvolvimento da arquitetura e componentes 3C.

O projeto *AulaNet* [5] está entre os exemplos onde o Modelo 3C de Colaboração foi utilizado como referência para o desenvolvimento de um sistema colaborativo ou *groupware*. O objetivo desse projeto foi construir um ambiente de aprendizagem colaborativa, tendo em vista a necessidade de propiciar aos alunos daquela universidade a experiência de trabalhar em grupo. O ambiente foi criado com base na colaboração, presente na interação entre aprendizes, colegas aprendizes, docentes e conteúdos didáticos.

Considerando que o ambiente *AulaNet* valoriza a abordagem colaborativa na dinâmica de criação de cursos online [5] e que um projeto de MOOC conectivista possui abordagem pedagógica semelhante a utilizada no *Aulanet*, esta dissertação de mestrado propõe utilizar o Modelo 3C de Colaboração como referência na condução do estudo exploratório para identificação de um conjunto de requisitos que atendam ao domínio MOOC conectivista.

O escopo deste estudo está relacionado à identificação de requisitos que deem suporte às práticas conectivistas de um cMOOC. Requisitos relacionados às demais propriedades

que caracterizam um MOOC, tais como participação aberta e massiva, não fazem parte do escopo desta pesquisa.

Com o conjunto de requisitos identificados a partir do Modelo 3C de Colaboração, será proposto um *framework* conceitual com o detalhamento necessário que guie a construção de ambientes de aprendizagem para MOOC conectivista. Com o *framework* conceitual, espera-se otimizar o processo de engenharia de requisitos para esse domínio, uma vez que novos ambientes poderão ser construídos a partir dos requisitos propostos pelo *framework* conceitual, reduzindo tempo e custo do projeto.

Os seguintes objetivos foram definidos para execução deste estudo, visando responder a questão de pesquisa e validar a solução proposta:

Objetivo Geral

- Identificar os requisitos do domínio MOOC que suportam a abordagem conectivista sob a ótica do Modelo 3C de Colaboração.

Objetivos Específicos

- Identificar na literatura os requisitos relacionados ao Modelo 3C para Sistemas Colaborativos.
- Identificar na literatura os requisitos relacionados à abordagem conectivista.
- Definir o conjunto de requisitos para o domínio MOOC conectivista com base no Modelo 3C de Colaboração.
- Avaliar os requisitos definidos para o domínio MOOC conectivista.

4. PROJETO DE AVALIAÇÃO DA SOLUÇÃO

A avaliação da solução proposta será realizada de acordo com as seguintes etapas:

- *Framework* conceitual para o domínio cMOOC: Criação de um *framework* conceitual com o detalhamento necessário sobre os requisitos que dão suporte às praticas conectivistas para construção de softwares do domínio cMOOC.
- Avaliação do *Framework* conceitual: avaliação do *Framework* conceitual e requisitos propostos, a ser realizada com base em opinião especializada formada por representantes da área de educação.
- Protótipo de software cMOOC: Construção de um protótipo de software a partir do *framework* conceitual para cMOOC. O protótipo será um ambiente de aprendizagem completo para execução de um curso.
- Estudo de caso: Realização de um estudo de caso utilizando-se um curso de curta duração da universidade a ser disponibilizado no ambiente de aprendizagem cMOOC. A participação de aprendizes será aberta ao público e os participantes serão acompanhados durante a execução do curso. Após encerramento, espera-se aplicar questionários e realizar entrevistas com participantes, bem como obter a opinião de especialistas em educação. Com base nos dados coletados, almeja-se avaliar o atendimento aos requisitos que dão suporte à abordagem conectivista do cMOOC propostos no *framework* conceitual.

5. ATIVIDADES JÁ REALIZADAS

Foi iniciado um mapeamento sistemático da literatura sobre MOOC que está sendo executado por um grupo de estudo formado por três alunos de mestrado da universidade, sob coordenação de seus orientadores. O planejamento e execução do mapeamento sistemático da literatura seguem as diretrizes propostas por Kitchenhan e Charters [8].

O objetivo principal do mapeamento é identificar características inerentes a MOOC. Os objetivos específicos são: construção de uma base conceitual para caracterização de MOOCs (tipos, definições, conceitos); mapeamento de experiências na utilização de MOOC ou na construção de plataformas e provedores MOOC; mapeamento de problemas enfrentados e soluções propostas; e possíveis tendências futuras de pesquisa.

Na primeira rodada de busca de artigos, por volta de 1.400 (mil e quatrocentos) estudos foram selecionados. Após aplicação dos critérios de inclusão e exclusão previstos no protocolo, em média 900 (novecentos) artigos foram mantidos e distribuídos entre o grupo de estudo para a etapa de extração de dados.

Os temas de pesquisa dos alunos foram definidos no início da fase de extração de dados, entre os quais, o tema desta dissertação. Embora as três pesquisas compartilhem uma parte comum do referencial teórico sobre MOOC, elas são independentes e suas respectivas contribuições não estão relacionadas.

Como o foco do mapeamento está restrito ao termo MOOC, foi necessário conduzir uma revisão da literatura para complementar o referencial teórico a ser utilizado nesta pesquisa. Foram selecionados artigos sobre Conectivismo, Sistemas Colaborativos, Modelo 3C de Colaboração, e requisitos para desenvolvimento de MOOC no contexto da engenharia de software. Os estudos selecionados são utilizados no delineamento do marco teórico e na justificativa que embasa a realização desta pesquisa.

6. CONCLUSÃO

Esta pesquisa propõe a realização de um estudo exploratório com objetivo de definir os requisitos, no contexto da engenharia de software, que atendam à abordagem utilizada em MOOCs projetados com base na Teoria Conectivista. A proposta prevê a utilização do Modelo 3C de Colaboração como referência para identificação desses requisitos. Com o resultado do estudo exploratório, espera-se criar um *framework* conceitual que apoie pesquisadores e profissionais da indústria no processo de engenharia de requisitos para implementação de softwares no domínio MOOC Conectivista (cMOOC).

7. AGRADECIMENTOS

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado da Bahia pelo apoio financeiro para realização deste estudo, através da concessão de bolsa de pesquisa BOL2509/2016.

8. REFERENCES

[1] M. H. Baturay. An overview of the world of moocs. *Procedia-Social and Behavioral Sciences*, 174:427–433, 2015.

[2] B. Dasarathy, K. Sullivan, D. C. Schmidt, D. H. Fisher, and A. Porter. The past, present, and future of

moocs and their relevance to software engineering. In *Proceedings of the on Future of Software Engineering*, pages 212–224. ACM, 2014.

[3] C. A. Ellis, S. J. Gibbs, and G. Rein. Groupware: some issues and experiences. *Communications of the ACM*, 34(1):39–58, 1991.

[4] A. Fassbinder, M. E. Delamaro, and E. F. Barbosa. Construção e uso de moocs: Uma revisão sistemática. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 25, page 332, 2014.

[5] H. Fuks, M. A. Gerosa, A. B. Raposo, and C. J. P. D. Lucena. O modelo de colaboração 3C no ambiente AulaNet. *Informática na Educação: Teoria & Prática*, 7(1):25–48, 2004.

[6] H. Fuks and M. Pimentel. *Sistemas colaborativos*. Elsevier Brasil, 2011.

[7] B. Grainger. Introduction to moocs: avalanche, illusion or augmentation. *URL: http://iite. unesco. org/pics/publications/en/files/3214722. pdf*, 2013.

[8] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering, 2007.

[9] M. Pimentel, M. A. Gerosa, D. Filippo, A. Raposo, H. Fuks, and C. J. P. Lucena. Modelo 3c de colaboração para o desenvolvimento de sistemas colaborativos. *Anais do III Simpósio Brasileiro de Sistemas Colaborativos*, pages 58–67, 2006.

[10] O. Rodriguez. The concept of openness behind c and x-moocs (massive open online courses). *Open Praxis*, 5(1):67–73, 2013.

[11] G. Siemens. Connectivism: A Learning Theory for the Digital Age. *International Journal of Instructional Technology and Distance Learning*, 1:1–8, 2005.

Arquitetura Publish/Subscribe baseada em semântica

Semantic-based Publish/Subscribe architecture

Antônio Pimenta Júnior
Programa de Pós-Graduação em Informática
Universidade Federal do Estado do Rio de Janeiro
Rio de Janeiro - RJ
antonio.junior@uniriotec.br

Leonardo Azevedo
Programa de Pós-Graduação em Informática
Universidade Federal do Estado do Rio de Janeiro
Rio de Janeiro - RJ
IBM Research
azevedo@uniriotec.br; lga@br.ibm.com

Flávia Santoro
Programa de Pós-Graduação em Informática
Universidade Federal do Estado do Rio de Janeiro
Rio de Janeiro - RJ
flavia.santoro@uniriotec.br

RESUMO

Integração e troca de informações entre sistemas é um problema desafiador em arquitetura de software. Heterogeneidade e a volatilidade das fontes de informação são inerentes no cenário de sistemas de diferentes fornecedores, construídos sobre diferentes filosofias e prioridades. *Publish/Subscribe* (*Pub/Sub*) é um padrão de integração de sistemas que pretende resolver esses problemas por ser capaz de determinar quando uma informação deve ser entregue. Propostas de soluções semânticas baseadas em ontologias buscam permitir um maior poder de representatividade e inferências. Porém, estas propostas possuem a limitação de exigir que sejam construídas as ontologias que descrevem as informações dos sistemas participantes. Este trabalho propõe uma arquitetura de integração alternativa baseada no padrão *Pub/Sub*. Através do uso de listas de termos para descrever as informações dos sistemas participantes, esta arquitetura permitirá um alto poder de inferência, sem a limitação de exigir a confecção de ontologias.

Palavras-Chave

Event-Driven Architecture, Semantic Publish/Subscribe

ABSTRACT

Integration and exchange of information between systems is a challenging problem in software architecture. Heterogeneity and volatility of information sources are inherent in the scenario of systems of different suppliers, built on different philosophies and priorities. Publish/Subscribe (Pub/Sub) is a system integration pattern that intends to solve these problems by being able to determine when an information should be delivered. Proposals for semantic solutions based on ontologies seek to allow a greater power of representativity and inferences. However, these proposals have the limitation of requiring that the ontologies that describe the information of the participating systems be constructed. This paper proposes an alternative integration architecture based on the Pub / Sub pattern. Through the use of term lists to describe the information of the participating systems, this architecture will allow a high power of inference, without the limitation of requiring the creation of ontologies.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•**Applied computing** → **Enterprise application integration; Information integration and interoperability;** •**Information systems** → *Mediators and data integration;*

Keywords

Event-Driven Architecture, Semantic Publish/Subscribe

1. INTRODUÇÃO

A integração e a troca de informações entre sistemas é um problema chave no mundo da tecnologia. Uma característica importante neste cenário é a heterogeneidade e a volatilidade das fontes de informação. Nesses cenários temos sistemas de diferentes fornecedores, construídos sobre diversas filosofias e prioridades [4].

Publish/Subscribe (*Pub/Sub*) é um padrão de integração de sistemas baseado na troca de eventos que tem ganhado notoriedade por ser considerado promissor para resolver esses desafios. Para isso, uma característica fundamental desse padrão é a capacidade de determinar quando um evento deve ser enviado ou não. Versões semânticas desse padrão, utilizando ontologias, vem sendo propostas na tentativa de permitir um alto poder de inferência. Porém, essas soluções possuem a limitação de exigir que sejam criadas ontologias para descreverem os eventos, o que muitas vezes não se justifica devido a complexidade desta tarefa [1, 2].

Este trabalho propõe uma arquitetura de integração baseada no padrão *Pub/Sub* que permita um alto poder de inferência, mas que não possua a limitação de exigir a criação de ontologias. Nesta abordagem, para a determinação das rotas de entrega, o servidor de eventos (mediador da comunicação) aplica, sobre as listas de termos que descrevem os eventos, processamento de linguagem natural e avaliação

de similaridade semântica associada à base de conhecimento linguístico *WordNet*.

O restante do artigo está organizado como segue. Na Seção 2 são apresentados os principais conceitos empregados. A Seção 3 detalha o problema. A Seção 4 apresenta a solução proposta. A Seção 5 apresenta o planejamento da avalização. Finalmente, a Seção 7 apresenta as conclusões.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Publish/Subscribe

Arquitetura Orientada a Eventos (*Event-driven Architecture* ou EDA) descreve um padrão de troca de mensagens de baixo acoplamento e eficiente no quesito trafego de informações. Ao contrário das propostas de integração mais tradicionais, em que a comunicação é ponto-a-ponto e síncrona, na arquitetura orientada a eventos os principais objetivos são garantir a flexibilidade de comunicação e o baixo acoplamento entre os sistemas [1].

O padrão *Publish/Subscribe* (*Pub/Sub*) é a especialização da EDA considerada mais promissora por ajudar a resolver desafios de integração [2]. Dentre os ganhos, destaca-se a redução do tráfego desnecessário de informações, uma vez que só ocorre transmissão de dados quando um novo evento é produzido.

Um *Evento* é uma mudança de estado de alguma informação relevante para o contexto do negócio que ocorre, por exemplo, devido à interação de um usuário com um sistema ou a algum processamento lógico que produziu ou alterou alguma informação. Como exemplo, podem ser citados o cadastro de uma nova informação, a ocorrência de uma compra com cartão de crédito, um saque, um sinal de um sensor, e o atingimento de uma cotação de moeda [1].

Os sistemas *Publishers* ou Publicadores são aqueles onde ocorre algum processamento lógico que produz ou altera alguma informação e que a disponibiliza como um *Evento*. Os sistemas *Subscribers* ou Assinantes são aqueles interessados em serem notificados da ocorrência de um evento no contexto do negócio. O *Events Server* ou Servidor de Eventos é um mediador da comunicação entre os publicadores e os assinantes. Ele é o responsável por receber as notificações dos publicadores, manter as assinaturas dos assinantes, armazenar temporariamente os eventos e os rotear/transmitir para os assinantes.

O padrão *Pub/Sub* permite que o assinante expresse seu interesse em determinados tipos ou padrões de eventos, para que, consequentemente, passe a ser notificado quando da ocorrência desses eventos no publicador. Geralmente, os sistemas assinantes estão interessados em receber eventos específicos. Existem diferentes tipos de implementações de *Pub/Sub* classificados de acordo com a forma de especificar os tipos de eventos de interesse do assinante: (i) Baseada na divisão dos canais de comunicação por assunto ou tópico; (ii) Baseado no conteúdo dos eventos; (iii) Baseado no tipo de dado que o evento encapsula. [3, 7, 11]

2.2 Similaridade Semântica

As medidas de similaridade semânticas podem ser utilizadas para determinar a similaridade de conceitos que muitas vezes não são lexicalmente similares. Para isto, estas medidas exploram relacionamentos linguísticos entre conceitos utilizando recursos externos, como a *WordNet*.

Lin e Sandkuhl [6] classificam as estratégias de medida de similaridade semântica em: (i) Baseadas em arestas (*edge-based*): a similaridade entre dois conceitos é determinada pela distância (caminho) entre os conceitos e a posição do conceito na taxonomia; (ii) Baseadas em informação (*information-based*): a similaridade é determinada considerando a quantidade de informação existente entre os conceitos em função das suas probabilidades de ocorrência em um corpus; e (iii) Híbridas: a similaridade é determinada combinando as duas abordagens anteriores.

Slimani [9] realizou um estudo comparativo entre os principais métodos de medida de similaridade, onde se destacaram as estratégias Wu e Palmer [10] e Lin [5] em relação à assertividade,.

3. PROBLEMA

A capacidade de expressar através de filtros quais são os eventos de interesse dos assinantes permite um melhor desempenho das soluções que utilizam o padrão (*Pub/Sub*). Contudo, as abordagens existentes são limitadas por não permitirem uma avaliação semântica das informações de interesse do serviço assinante em relação às informações publicadas pelos serviços fornecedores [11]. Essas abordagens são inadequadas em ambientes heterogêneos e dinâmicos, onde o publicador e o assinante podem ter esquemas de eventos diferentes estruturalmente. Nessas soluções, eventualmente algum evento pode ser entregue a um assinante que não tem interesse naquele tipo de informação, o que representa um tráfego desnecessário na rede. De outra forma, algum evento que possui informações importantes pode ser filtrado e deixar de ser entregue para o assinante [2] [11].

Na literatura existem alguns trabalhos que propõem implementações que tentam resolver estes problemas através da adoção de ontologias para permitir a análise semântica dos eventos. Em Skovronski *et al.* [8] propõem uma arquitetura *Pub/Sub* semântica baseada em ontologias. Os sistemas publicadores devem informar a ontologia que descreve o conteúdo que irão publicar. Desta forma, todos os eventos daquele publicador devem ser expressos através de linguagem XML, onde as *tags* são nomeadas com classes ou propriedades da ontologia. Os assinantes assinam as publicações de determinada ontologia. Todo o processamento de filtro dos eventos de interesse é feito no contexto do assinante. Para isso a solução propõe o uso de SPARQL como linguagem de subscrição.

Embora a solução usando ontologias tenha alcançado bons resultados, este tipo de abordagem exige que sejam confeccionadas as ontologias para a descrição dos eventos publicados, o que é extremamente complexo. Além disso, esta solução exige que cada serviço assinante tenha conhecimento sobre cada ontologia de cada evento de interesse, provocando um *overhead* no sistema assinante para processar o filtro dos eventos de interesse através de SPARQL.

Sendo assim, levando em consideração que as situações reais do mundo corporativo possuem nível relativamente baixo de complexidade de informações tratadas, o ganho na adoção da solução não se justificaria visto o alto custo de gerar e tratar as ontologias.

4. SOLUÇÃO

Este trabalho propões uma arquitetura que permita maior expressividade e flexibilidade aos filtros, para que, mesmo

em situações onde exista heterogeneidade entre o publicador e o assinante, o servidor de eventos seja capaz de determinar se um evento deve ou não ser entregue a um determinado assinante sem comprometer o desempenho dos sistemas.

A proposta baseia-se na utilização de termos para descrever os eventos dos fornecedores e o interesse dos assinantes. Para determinar se um evento deve ou não ser enviado a um assinante, as listas de termos do publicador e do assinante serão submetidos para avaliação de uma rotina de verificação de similaridade semântica. Com o resultado determina-se se o evento possui alguma informação de interesse do assinante.

Para confrontar as listas de termos e determinar a similaridade, serão utilizados os algoritmos de verificação de similaridade semântica destacados em Slimani [9], que utilizam como base de conhecimento a *WordNet*. No trabalho de Slimani os algoritmos de Lin e Wu e Palmer são destacados devido sua elevada eficiência. Por esse motivo, neste trabalho iremos considerar os resultados desses dois algoritmos. Na nossa solução, é dada a liberdade para que os publicadores e os assinantes criem as listas de termos que descrevem seus eventos e necessidade nos termos.

Para se conectar ao servidor de eventos, os Publicadores e Assinantes devem submeter sua lista de termos. Dessa forma, associada a cada serviço existirá uma lista de termos que descreve o conteúdo publicado ou a necessidade de informação, dependendo do tipo de serviço.

Quando um novo Assinante subscreve ao servidor de eventos será realizado um confrontamento entre os termos de sua lista e os termos dos Publicadores já conectados. A rotina de verificação irá analisar cada combinação e determinar a similaridade. Com o resultado dessa verificação é montada a tabela de roteamento, a qual determina as rotas de entrega de eventos entre os publicadores e os assinantes.

O servidor de eventos utilizará a tabela de roteamento como base para seu funcionamento, de forma que quando ocorre alguma notificação de evento por parte dos publicadores, a tabela de roteamento é verificada para determinar quais assinantes devem ser notificados.

5. AVALIAÇÃO DA PROPOSTA

Para avaliar a proposta, será implementada uma solução de referência para o modelo arquitetural proposto. Esta solução será submetida a experimentos simulando cenários reais. Serão coletados dados sobre a eficiência da proposta em relação à sua capacidade de inferir se um evento deve ser entregue a um assinante, de acordo com o especificação do conteúdo do evento e do interesse do assinante. Nesta análise serão avaliadas a Precisão [1], a Cobertura[2] e a Medida-F[3].

Com objetivo de validar os resultados da abordagem proposta e da solução de referência, será feita uma comparação com os resultados do trabalho de Skovronski *et al.* [8].

6. CONCLUSÃO

O padrão *Publish/Subscribe* vem ganhando notoriedade por ser considerado a especialização da *Event-driven Architecture* mais promissora por ajudar a resolver os principais desafios de integração de sistemas.

A capacidade de expressar através de filtros quais são os tipos de eventos de interesse dos assinantes permite um melhor desempenho das soluções que utilizam este paradigma. Contudo, a maioria das abordagens existentes são limitadas por não permitirem uma avaliação semântica das informações de interesse do serviço assinante em relação às informações publicadas pelos serviços fornecedores.

Este trabalho propôs uma nova abordagem automatizada de entrega de eventos baseada na semântica do conteúdo para as arquiteturas *Pub/Sub*. Nesta abordagem, listas de termos são utilizadas para descrever os eventos dos publicadores e o interesse dos assinantes. Algoritmos de análise de similaridade semântica são utilizados para determinar a relação entre os eventos produzidos pelos publicadores e o interesse dos assinantes. Com o resultado dessa análise o servidor de eventos é capaz de determinar com mais precisão o encaminhamento dos eventos.

7. REFERENCES

[1] K. Chandy and W. Schulte. *Event Processing: Designing IT Systems for Agile Companies*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 2010.

[2] M. Diallo, V. Sourlas, P. Flegkas, S. Fdida, and L. Tassiulas. A content-based publish/subscribe framework for large-scale content delivery. *Computer Networks*, 57(4):924 – 943, 2013.

[3] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec. The many faces of publish/subscribe. *ACM Comput. Surv.*, 35(2):114–131, June 2003.

[4] S. Guo. *Using Semantic Mapping for Semantic Based Publish/Subscribe System*. PhD thesis, School of Computer Science and Statistics, Trinity College, 2009.

[5] D. Lin. Principle-based parsing without overgeneration. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 112–120. Association for Computational Linguistics, 1993.

[6] F. Lin and K. Sandkuhl. A survey of exploiting wordnet in ontology matching. In *IFIP International Conference on Artificial Intelligence in Theory and Practice*, pages 341–350. Springer, 2008.

[7] S. Microsystems. *Sun Java System Message Queue 4.1 Developer's Guide for Java Clients*. 2007.

[8] J. Skovronski and K. Chiu. Ontology based publish subscribe framework. In *iiWAS'2006 - The Eighth International Conference on Information Integration and Web-based Applications Services, 4-6 December 2006, Yogyakarta, Indonesia*, pages 49–58, 2006.

[9] T. Slimani. Description and evaluation of semantic similarity measures approaches. *International Journal of Computer Applications*, 80:25–33, 2013.

[10] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[11] L. Zeng and H. Lei. A semantic publish/subscribe system. In *E-Commerce Technology for Dynamic E-Business, 2004. IEEE International Conference on*, pages 32–39. IEEE, 2004.

[1]Taxa de similaridades corretamente encontradas sobre o número total de similaridades retornadas.

[2]Taxa de similaridades corretamente encontradas sobre o número total de similaridades esperadas.

[3]Medida harmônica entre a precisão e a cobertura.

Método de Extração de Informação Baseado em Ontologias para Accountability das OSCIPs de Arte e Cultura do Rio de Janeiro

Alternative Title: Information Extraction Method Based on Ontologies for Accountability of OSCIPs of Art and Culture of Rio de Janeiro

Patrick Ferreira Barroso
Programa de Pós-graduação em Informática
Universidade Federal do Estado do Rio de Janeiro
Avenida Pasteur 458, Urca
Rio de Janeiro, RJ, Brasil
patrick.barroso@uniriotec.br

Renata Araujo
Programa de Pós-graduação em Informática
Universidade Federal do Estado do Rio de Janeiro
Avenida Pasteur 458, Urca
Rio de Janeiro, RJ, Brasil
renata.araujo@uniriotec.br

RESUMO

As OSCIPs (Organizações da Sociedade Civil de Interesse Público) de Arte e Cultura do RJ são instituições não-governamentais que executam Políticas Culturais através de parcerias realizadas com o Estado, realizando ações de interesse público. Elas possuem o dever de prestar contas de suas ações à sociedade, conforme determinação da Lei de Acesso à Informação (12.527/2011) e a Lei das OSCIPs (9.790/1999). O trabalho apresenta: (i) o baixo atendimento aos requisitos legais das informações organizacionais e operacionais das OSCIPs; (ii) a dispersão das suas informações na web, que dificulta a manipulação dos dados e acesso à informação; (iii) proposta de solução utilizando o método de Extração da Informação Baseado em Ontologias (EIBO).

Palavras-chave

Políticas Públicas, Políticas Culturais, Accountability, Terceiro Setor, Extração da Informação, Ontologias.

ABSTRACT

The OSCIPs (Civil Society Organizations of Public Interest) of Art and Culture of RJ are non-governmental institutions that execute Cultural Policies through partnerships with the State, performing actions of public interest. They have a duty to account for their actions to society, as determined by the Law on Access to Information (12.527/2011) and the OSCIPs Law (9.790/1999). The work presents: (i) low compliance with the legal requirements of OSCIPs organizational and operational information on the web; (ii) the dispersion of your information, which makes it difficult to manipulate the data and access to information; (iii) solution proposal using Ontology-Based Information Extraction (EIBO).

SBSI 2017, June 5th–8th, 2017, Lavras, Minas Gerais, Brazil.

CCS Concepts

• **Information systems→Information retrieval→Information Extraction** • **Social and professional topics→Computing / technology policy→Government technology policy→Government regulations**

Keywords

Public Policies, Cultural Policies, Third Sector, Accountability, Information Extraction, Ontologies.

1. INTRODUÇÃO

Na realização de Políticas Públicas, o Governo Federal utiliza diversos meios para atingir seus objetivos. Para isso, conta com a ajuda de diversas instituições, dentre elas as não governamentais (ONGs), que fazem parte do Terceiro Setor.

O Terceiro Setor desempenha o papel de supridor das necessidades de uma parcela da sociedade, carente de assistência social, educacional e cultural, não satisfeitas pelo Estado, através de organismos criados para esse fim por grupos empresariais (Fundações) ou por pessoas que se reúnem para esse fim em defesa de um mesmo ideal (entidades sem fins lucrativos) [1].

Accountability pode ser entendido como transparência dos governantes na prestação de contas e a responsabilização destes mesmos governantes pelos seus atos [15]. No âmbito do Terceiro Setor, [3] informa que é o ato de prestar contas de forma responsável nas organizações sem fins lucrativos e demonstrar que cumpriu a sua missão, ou seja, que utilizou corretamente os recursos recebidos de doações.

A Lei de Acesso à Informação (LAI, 12.527/2011) determina diretrizes tanto ao Governo quanto para Instituições Não Governamentais, para que façam a devida prestação de contas (*accountability*) de suas ações por meios digitais, pois são de interesse público. A LAI determina que "é dever dos órgãos e entidades públicas promover, independentemente de requerimentos, a divulgação em local de fácil acesso, no âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas."

As OSCIPs (Organizações da Sociedade Civil de Interesse Público) são uma das categorias de instituições não-governamentais do Terceiro Setor, que possuem a prerrogativa de realizar convênios e parcerias com o Estado, a fim de promover ações de interesse público. A Lei das OSCIPs (9.790/1999) exige que tais instituições "prestem contas de todos os recursos e bens de origem pública", conforme determinações constitucionais. [17] informa que a *accountability* das OSCIPs não é apenas uma imposição legislativa, mas o compromisso de prestar contas com aqueles que financiam essas organizações.

2. O PROBLEMA

Foi realizado um levantamento manual em sites de busca para avaliar se as OSCIPs de Arte e Cultura do Rio de Janeiro divulgam as informações organizacionais e operacionais determinadas por lei. Para isso, foram realizados os seguintes passos: (i) levantamento das OSCIPs de Arte e Cultura do RJ no site do Ministério da Justiça; (ii) elaboração de um arcabouço de itens de informação determinados pela Lei de Acesso à Informação (LAI) e Lei das OSCIPs; (iii) mapeamento manual das informações em ferramentas de busca na web; (iv) análise dos dados coletados.

Para realizar o levantamento das OSCIPs de Arte e Cultura da cidade do Rio de Janeiro, foi feita uma consulta através do portal do Ministério da Justiça [9], que é o órgão responsável por qualificá-las. Foram encontradas dezenove (19) instituições (Tabela 1).

Tabela 1. Relação das OSCIPs de Arte e Cultura RJ

CNPJ	NOME DA INSTITUIÇÃO (OSCIP)
05596539000179	INSTITUTO DE DESENVOLVIMENTO, ESTUDO E INTEGRAÇÃO PELA ANIMAÇÃO - IDEIA
31886799000199	ASSOCIAÇÃO CASA DO PONTAL - COLEÇÃO JACQUES VAN DE BEUQUE
07379466000199	ASSOCIAÇÃO CULTURAL BURITI - A C B
05977454000130	ASSOCIAÇÃO DE CULTURA E MEIO AMBIENTE - ACMA
03360608000115	ASSOCIAÇÃO DOS AMIGOS DA ARTE POPULAR BRASILEIRA
06370226000160	ASSOCIAÇÃO O ECO - O ECO
06555811000135	ASSOCIAÇÃO VIA CULTURAL BRASIL - VIA BRASIL
02979479000185	CÂMARA DE CULTURA, COMÉRCIO E TURISMO BRASIL - PAÍSES AFRICANOS
01315883000191	CEMCO - CENTRO DE ESTUDOS DE MÚSICA CORAL
04906029000198	CENTRO PRESERV DE PROMOÇÃO DO DESENVOLVIMENTO SUSTENTADO - PRESERV
06085782000195	CINEMA NOSSO
05075576000131	FUNDAÇÃO CULTURAL ARO
04723294000130	INSTITUIÇÃO SOCIAL CULTURAL ALEGRIA DE LER - ISCAL
07682216000123	INSTITUTO ART DÉCO BRASIL
07508414000175	INSTITUTO BRASILEIRO DE GESTÃO PÚBLICA - IBGP
03808720000176	INSTITUTO DE IMAGEM E CIDADANIA RIO DE JANEIRO
04521945000100	INSTITUTO EMBRATEL CLARO
05381947000103	INSTITUTO PÉ NO CHÃO - IPC
07579027000120	INSTITUTO TAMANDUÁ SYNAPSE CULTURAL - INSTITUTO

Para a criação do arcabouço dos itens de informação, foi extraído da LAI e Lei das OSCIPs (Tabela 2). O grupo Organizacional reflete a necessidade de apresentar os dados referentes à própria instituição (estrutura, metas, objetivos e projetos). Para o grupo Contratual, são considerados itens referentes ao Termo de Parceria, que são considerados obrigatórios quando a OSCIP em questão possuir convênio em andamento com alguma instituição governamental.

Para o mapeamento dos itens de informação, foi realizada uma pesquisa manual, mediante ferramenta de busca na web, para verificar se cada OSCIP possui as informações requeridas pelo arcabouço de referência (Tabela 2).

No eixo vertical da Tabela 3 consta a relação das OSCIPs de Arte e Cultura do RJ e no eixo horizontal os itens de informação (referentes à Tabela 2). Um visto foi colocado no quadrado correspondente ao item caso este fosse atendido. Em caso de não atendido, o quadrado ficaria em branco.

Tabela 2. Arcabouço de itens de informações das OSCIPs exigidas por lei

ITEM DE INFORMAÇÃO	GRUPO	DISPOSITIVO LEGAL
Estrutura organizacional (competências, endereços, telefones e horários de atendimento)	ORGANIZACIONAL	LAI (Art. 8o / § 1o / I)
Missão		LAI (Art. 7o V)
Metas e objetivos		LAI (Art. 7o VII a)
Serviços Prestados		LAI (Art 7o X)
Projetos Sociais		LAI (Art. 8o / § 1o / V)
Objeto do Projeto ou Programa	CONTRATUAL	LEI OSCIP (Art. 10 § 1o)
Metas e objetivos		LAI (Art. 7o VII a), LEI OSCIP (Art. 10 § 2o I e II)
Cronograma de atividades		LEI OSCIP (Art. 10 § 1o), LEI OSCIP (Art. 10 § 2o II)
Critérios de Avaliação (indicadores)		LAI (Art. 7o VII a), LEI OSCIP (Art. 10 § 2o III)
Previsão de receita e despesas		LAI (Art. 8o / § 1o / III), LEI OSCIP (Art. 10 § 2o IV)
Extrato Demonstrativo de Execução		LEI OSCIP (Art. 15B II)
Relatório Anual de Execução de Atividades		LEI OSCIP (Art. 15B I)
Demonstração de Origem de Aplicação de Recursos		LEI OSCIP (Art. 15B VI)
Demonstração do déficit ou superávit do exercício;		LAI (Art. 7o VII b)

Em suma, foi concluído que:

- Apenas 11% das OSCIPs atenderam a todos os itens de informação e nenhuma atendeu completamente ao grupo Contratual;
- Atendimento de 49% para o grupo Organizacional e 46% no geral das OSCIPs em questão.
- As informações de cada OSCIP estão dispersas na internet e de forma desestruturada. Através da pesquisa manual realizada, foi verificado que as informações constam em fontes diversas, dentre as principais: websites próprios, portal ONGs Brasil [13], portal governamental de convênios SICONV [18], redes sociais de pessoas físicas ou jurídicas, blogs, dentre outros.

Tabela 3. Resultado do mapeamento dos itens de informação das OSCIPs na web

MAPEAMENTO DOS ITENS DE INFORMAÇÃO PARA OSCIPS ARTE E CULTURA RJ	ORGANIZACIONAL				CONTRATUAL									
	1. Estrutura Organizacional	2. Missão	3. Metas e Objetivos	4. Serviços Prestados	5. Projetos Sociais	1. Objeto do Projeto ou Programa	2. Metas e Objetivos	3. Cronograma de atividades	4. Critérios de avaliação (indicadores)	5. Previsão de receita e despesas	6. Extrato Demonstrativo de Execução	7. Relatório Anual de Atividades	8. Demonstração do origem e aplicação de recursos	9. Demonstração do déficits ou superavit do exercício
ASSOCIAÇÃO CASA DO PONTAL - COLEÇÃO JACQUES VAN DE BEUQUE		✔		✔	✔	SEM CONVÊNIOS (NÃO SE APLICA)								
ASSOCIAÇÃO CULTURAL BURITI - A C B				✔										
ASSOCIAÇÃO DE CULTURA E MEIO AMBIENTE - ACMA		✔		✔	✔									
ASSOCIAÇÃO DOS AMIGOS DA ARTE POPULAR BRASILEIRA				✔		✔	✔					✔		
ASSOCIAÇÃO O ECO - O ECO	✔	✔		✔	✔	SEM CONVÊNIOS (NÃO SE APLICA)								
ASSOCIAÇÃO VIA CULTURAL BRASIL - VIA BRASIL				✔										
CÂMARA DE CULTURA, COMÉRCIO E TURISMO BRASIL - PAÍSES AFRICANOS				✔										
CEMCO - CENTRO DE ESTUDOS DE MÚSICA CORAL				✔										
CENTRO PRESERV DE PROMOÇÃO DO DESENVOLVIMENTO SUSTENTADO - PRESERV		✔	✔	✔										
CINEMA NOSSO		✔		✔	✔									
FUNDAÇÃO CULTURAL ARO		✔		✔	✔									
INSTITUIÇÃO SOCIAL CULTURAL ALEGRIA DE LER - ISCAL				✔										
INSTITUTO ART DÉCO BRASIL	✔	✔	✔	✔	✔									
INSTITUTO BRASILEIRO DE GESTÃO PÚBLICA - IBGP		✔	✔	✔	✔									
INSTITUTO DE DESENVOLVIMENTO, ESTUDO E INTEGRAÇÃO PELA ANIMAÇÃO - IDEIA				✔										
INSTITUTO DE IMAGEM E CIDADANIA RIO DE JANEIRO		✔		✔	✔	✔	✔					✔		
INSTITUTO EMBRATEL CLARO	✔	✔	✔	✔	✔	SEM CONVÊNIOS (NÃO SE APLICA)								
INSTITUTO PÉ NO CHÃO - IPC				✔										
INSTITUTO TAMANDUÁ SYNAPSE CULTURAL - INSTITUTO		✔		✔	✔	✔	✔					✔		

A ameaça à validade a ser considerada é que a busca foi realizada de forma manual e sem aplicação de métodos científicos, então não podem ser constatadas as informações de forma precisa e confiável.

3. PROPOSTA DE SOLUÇÃO

Para o tratamento dos problemas da dispersão de dados e a necessidade de um método sistemático de avaliação dos itens de informação, a proposta é aplicar o método de Extração da Informação Baseado em Ontologias (EIBO) adaptando o método apresentado por [5], [12] e [20].

Extração (ou recuperação) de informação é a tarefa de extrair informação de forma automática a partir de documentos legíveis por computador. Essa extração pode ser realizada por meio de métodos puramente matemáticos (estatísticos) ou pela utilização de métodos e técnicas de Processamento de Linguagem Natural (PLN) [6]. O principal objetivo de um Sistema de Extração de Informação (SEI) é identificar fragmentos de texto em um conjunto de documentos (*corpus*) que preencham corretamente campos de informação (*slot*) definidos previamente. Esses campos formam

uma saída estruturada (*template*) que determina quais informações deverão ser encontradas [12].

Uma ontologia é uma estrutura conceitual que visa representar formalmente os conceitos e suas relações, regras e restrições lógicas de um determinado domínio, e pode ser definida por meio de linguagens processáveis por computadores [10]. Um Sistema de Extração de Informação Baseado em Ontologias (SEIBO) é um sistema que processa textos oriundos de uma fonte de dados não estruturada ou semiestruturada, através de um mecanismo guiado por ontologias para extrair certos tipos de informações, além de apresentar sua saída usando ontologias [20]. As ontologias servem como ferramentas para a seleção dos termos que irão compor a consulta do usuário [10] e facilita o processo de interpretação dos dados pelas ferramentas de recuperação [16].

Criação do método de EIBO (conforme Figura 1) seguirá os seguintes passos principais:

1. Elementos de busca: (i) Definição do corpus, que são documentos semiestruturados ou não estruturados na web, coletado através de *crawlers* que possuem a função de pesquisar e indexar os documentos baseados nos itens de informação de referência (Figura 1). (ii) Modelagem do *template* necessário para definição da ontologia inicial e; (iii) slot de entrada, que são os termos e expressões regulares necessárias para a extração.
2. Processo de extração (classificação de candidatos à instâncias): Serão utilizadas ferramentas de Processamento de Linguagem Natural (PLN), como tokenização, divisão de sentenças, filtros semânticos, etiquetagem e sintagmas nominais, entre outros.
3. Povoamento de Ontologias: Definição dos candidatos às instancias e o carregamento dos dados, gerando como resultado a "ontologia povoada", ou seja, os dados gerados no formato aberto que possibilitará a manipulação do usuário.

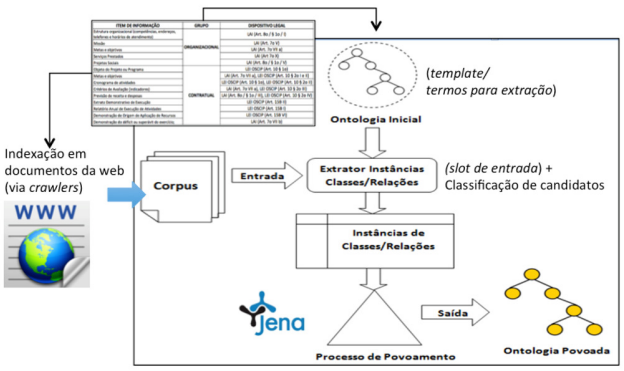


Figura 1. Modelo EIBO adaptado [12]

A linguagem de programação a ser utilizada para aplicação do método EIBO será o Java [7], a API de povoamento de ontologias será o Jena [8], plataforma de desenvolvimento Eclipe [4] e o processamento de linguagem natural (PLN) será o Open PNL [14].

4. METOLOGIA DE PESQUISA

A pesquisa utilizará o método DSR (*Design Science Research*) através de uma abordagem empírica quantitativa e qualitativa exploratória. De acordo com [19], DSR é um paradigma epistemológico para construir conhecimento sobre o mundo a partir do projeto de artefatos. As pesquisas em DSR se estruturam em ciclos interligados, um sobre o design do artefato e outro sobre o conhecimento científico que fundamenta o design.

Serão desenvolvidos dois artefatos para esse estudo: o primeiro é o método EIBO, e posteriormente um Portal de Dados Abertos das OSCIPs de Arte e Cultura do Rio de Janeiro, onde os dados extraídos (ontologias povoadas) serão disponibilizados para consulta e manipulação do cidadão.

5. AVALIAÇÃO

Serão aplicados dois ciclos com a Metodologia DSR. Para o primeiro ciclo, será realizado um experimento para avaliar o modelo EIBO desenvolvido, utilizando métodos matemáticos de Precisão, Cobertura e Medida-F [2], [20], [21], e a determinação dos candidatos à instância através do modelo MCC (Medida de Confiança Combinada) [12].

No segundo ciclo será desenvolvido um portal de dados abertos das OSCIPs em questão, onde as instâncias carregadas do método EIBO serão disponibilizadas em diferentes formatos para consulta e manipulação do usuário (cidadão). Será realizado um estudo de caso para avaliar dados quantitativos e qualitativos de *accountability* junto ao cidadão, mediante questionários e entrevistas.

6. CONCLUSÃO

Foi exposta a problemática da dispersão de dados das OSCIPs de Arte e Cultura do Rio de Janeiro na web e a falta de um modelo confiável de extração a fim de avaliar se tais instituições atendem às informações exigidas por lei (LAI e Lei das OSCIPs).

Espera-se, com a proposta de solução do método EIBO e a construção de um portal de dados abertos, que as OSCIPs promovam a devida *accountability* de suas ações públicas ao cidadão, promovendo acesso aos dados de forma única e estruturada, apontando os itens de informação atendidos ou não, de acordo com os dispositivos legais.

As organizações devem ser transparentes e prestar contas aos seus públicos de interesse, projetando uma imagem favorável da organização ou fazer esforços de legitimação para evitar os malefícios da perda da legitimidade, os conflitos e a diminuição do apoio social, pois o objetivo da organização é ser legítima [11].

7. REFERÊNCIAS

[1] Brito, P. P.; Oliveira, M. C.; Santos, S. M. dos; Oliveira, B. C. de. (2008) A utilização dos demonstrativos contábeis como instrumento de apoio a gestão nas organizações não governamentais: um estudo de caso no estado do Ceará. Revista Alcance, 15(1), 61 – 80.

[2] Carlson, A., Betteridge, J., Wang, R. C. Semi-supervised learning for information extraction. In Proceedings of the 3nd ACM International Conference on Web Search and Data Mining (WSDM), pp. 101- 110. 2010

[3] Carneiro, A., Oliveira, D., Torres, L. Accountability e Prestação de Contas das Organizações do Terceiro Setor: Uma Abordagem à Relevância da Contabilidade. Revista Sociedade, Contabilidade e Gestão. 2011.

[4] Eclipe Neon. https://eclipse.org/downloads/packages/eclipse-ide-java-developers/neon1a. Acesso em: 20/03/2017

[5] Ferneda, E. Introdução Aos Modelos Computacionais de Recuperação de Informação. Editora Ciência Moderna. 2012.

[6] Grishman, R. Information extraction; techniques and challenges. In: INTERNATIONAL SUMMER SCHOOL SCIE, 1997, New York.

[7] Java. http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html. Acesso em: 20/03/2017

[8] Jena 3.1 API for Java. https://jena.apache.org/download/index.cgi . Acesso em: 20/03/2016

[9] Ministério da Justiça. http://justica.gov.br/acesso-a-sistemas/consulta-a-entidades-qualificadas . Acesso em: 04/11/2016

[10] Neto, J. J., Ferneda, E. Ontologia como Recurso de Padronização Terminológica no Processo de Recuperação da Informação. Inf. Pauta. UNESP. 2016

[11] O'Donovan, G. (2002) Environmental disclosures in annual report. Extending the applicability & predictive power of legitimacy theory. Accounting, Auditing & Accountability Journal. 15(3), 344- 371.

[12] Oliveira, H. T. A. Um Método Não Supervisionado para o Povoamento de Ontologias na Web. Universidade Federal de Pernambuco. UFPE. 2013.

[13] ONGs Brasil. http://www.ongsbrasil.com.br . Acesso em 19/11/2016

[14] OPEN PNL. https://sourceforge.net/projects/openpnl . Acesso em: 01/05/2017

[15] Pinho, J. A. Gomes de. Raupp, F. M. Construindo a Accountability em portais eletrônicos de câmaras municipais: um estudo de caso em Santa Catarina. Cadernos Ebape FGV, Rio de Janeiro, v. 9, nº 1, artigo7. 2011.

[16] Santarem Segundo, J. E. Web Semântica, Dados ligados e Dados Abertos: uma visão dos desafios do Brasil frente às iniciativas internacionais. XVI Encontro Nacional de Pesquisa em Ciência da Informação (XVI ENANCIB). 2015.

[17] Santos, R. Origens De Recursos Das Organizações Da Sociedade Civil De Interesse Público (Oscip): uma abordagem à luz da accountability e das teorias stakeholders e legitimidade.

[18] SICONV. https://www.convenios.gov.br . Acesso em 19/11/2016

[19] Silva, Adilson R. Gamificação e Inteligência Coletiva para Promover a Participação em Sistema de Bate-Papo para Educação. UNIRIO (2016)

[20] Wimalasuriya, D., Dou, D. Ontology-based information extraction: an introduction and a survey of current approaches, In Journal of Information Science (JIS), vol. 36, issue 3, pp. 306-323, 2010

[21] Yildiz, B. Ontology-Driven Information Extraction. PHD These, Vienna University of Technology, 2007.

Seleção de canal para reconhecimento biométrico baseado em sinais de EEG

Channel selection for EEG-based biometric recognition

Rodrigo A. de Freitas Vieira
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
São Paulo, Brasil
rodrigo.vieira@usp.br

Clodoaldo A. de Moraes Lima
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
São Paulo, Brasil
c.lima@usp.br

RESUMO

A identificação biométrica de pessoas é de considerável importância para segurança. Biometria biomédica tem sido ativamente investigada apenas na última década. Embora a especificidade para os indivíduos tenha sido observada há algumas décadas, o processo de aquisição mais complicado e o tempo de espera impediram sua aplicação em controle de acesso. O reconhecimento biométrico baseado em sinais de eletroencefalograma (EEG) é reconhecido como um dos mais resistentes à fraudes, porém ainda é considerado inviável para aplicações práticas. Este projeto visa explorar o problema de seleção de canais de EEG para o reconhecimento biométrico. A abordagem proposta empregará técnicas de aprendizado de máquina para extração características dos sinais de EEG e para a seleção dos canais mais relevantes visando o reconhecimento biométrico. Espera-se que este trabalho ajude na evolução da pesquisa para aplicações práticas do reconhecimento biométrico baseada em sinais de EEG.

Palavras-Chave

Sistemas Biométricos, identificação, eletroencefalograma, seleção de canais

ABSTRACT

Biometric identification of person plays an important role for security. Biomedical biometrics has been actively investigated only in the last decade. Although specificity for individuals was observed a few decades ago, the more complicated acquisition process and waiting time prevented its application in access control. The biometric recognition based on electroencephalogram (EEG) signals is one of the most fraud resistant, however, it is still considered unfeasible for practical applications. The proposed approach aims to use machine learning techniques for feature extraction of the EEG signals, and to select the most relevant channels for biometric recognition. It is hoped that this work will help in the evolution of the research for practical applications of biometric recognition based on EEG signals.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•Information systems → Information systems applications; •Computing methodologies → Biometrics; Machine learning;

Keywords

Biometric systems, identification, electroencephalogram, channel selection

1. INTRODUÇÃO

Sistemas de reconhecimento biométrico são métodos para identificação ou autenticação de pessoas baseado em características físicas ou comportamentais intrínsecas ao indivíduo [4]. Com o aumento na importância da segurança, sistemas de reconhecimento biométrico se tornaram cada vez mais comuns em aplicações práticas. Modalidades biométricas tais como íris, impressão digital e face, já foram bastante estudadas e exploradas.

Nos últimos anos, o reconhecimento biométrico sofreu avanços significativos em termos de confiabilidade e precisão, e essas modalidades biométricas têm alcançado um bom desempenho em aplicações práticas. No entanto, mesmos os sistemas biométricos mais avançados ainda enfrentam alguns problemas de aplicabilidade ao mundo real. Dessa forma, busca-se o desenvolvimento de sistemas biométricos mais robustos e com alta resistência a fraudes.

Recentemente, aumentou-se o interesse em usar sinais elétricos cerebrais como característica fisiológica para a próxima geração de sistemas biométricos [5]. A utilização de sinais elétricos cerebrais na biometria é promissora, tendo como principal vantagem ser uma das técnicas mais resistentes a fraudes [10]. Esses sinais são dados biomédicos e integram uma categoria de novas modalidades de reconhecimento biométrico que engloba sinais tipicamente utilizados em diagnósticos clínicos.

Há diferentes métodos para identificação e captura de sinais elétricos cerebrais tais como imagem por ressonância magnética funcional, tomografia por emissão de pósitrons, magnetoencefalografia e eletroencefalografia. Dentre esses métodos a eletroencefalografia (EEG) é uma escolha interessante para biometria biomédica em função de sua a alta

resolução temporal, baixo custo, portabilidade e fácil montagem para exames quando comparado com as outras técnicas [10].

Diferentemente das modalidades biométricas tradicionais tais como a face, a íris ou impressão digital, os sinais de EEG não são expostos, possuindo uma natureza "secreta", o que atribui ao sistema de reconhecimento biométrico um alto nível de privacidade, quando comparado a outras técnicas [3]. Apesar das vantagens relacionadas ao reconhecimento biométrico baseado em EEG, pesquisas ainda são necessárias para o desenvolvimento de técnicas que permitam a utilização prática dessa tecnologia.

O objetivo deste projeto é investigar o problema de seleção de canais para biometria baseada em EEG propondo diferentes abordagens para a seleção dos canais de EEG mais representativos para reconhecimento biométrico. Este artigo é organizado como segue: a Seção 2 apresenta o problema de seleção de canais; a Seção 3 apresenta a proposta do projeto; a Seção 4 descreve as formas de avaliação dos métodos propostos; a Seção 5 descreve as atividades já realizadas no projeto; e a Seção 6 apresenta as conclusões.

2. APRESENTAÇÃO DO PROBLEMA

A dificuldade de se realizar a captura do sinal de EEG é um obstáculo para o desenvolvimento de sistemas práticos. O procedimento de obtenção de um EEG é inconveniente para o indivíduo, requerendo a utilização de um capacete de EEG, com eletrodos envoltos em gel condutivo [1]. Com essas condições, a aplicação prática da biometria baseada em EEG é inviável. Novas tecnologias estão sendo desenvolvidas a fim de contornar essas dificuldades e facilitar a obtenção desses sinais. Essas tecnologias utilizam eletrodos secos e examinam uma quantidade reduzida de canais.

A diminuição na captura de canais de EEG vem ao encontro de uma questão importante para classificação, o problema de dimensionalidade. O emprego de uma análise de EEG com uma grande quantidade de canais implica em um custo computacional elevado para técnicas de aprendizado de máquina e, consequentemente, um maior custo computacional. A figura 1 apresenta uma configuração do Sistema Internacional 10-10 para captura de sinais de EEG com 64 eletrodos.

Uma solução já explorada em outras áreas de pesquisa, mas ainda pouco aplicada na biometria, é a seleção prévia de canais mais relevantes de EEG para otimização da tarefa. Reduzir a quantidade de canais de EEG mantendo-se altas taxas de identificação é crucial para o uso efetivo de EEG em aplicações biométricas [8], isto porque alguns canais carregam informações importantes, enquanto outros prejudicam ou não influenciam na identificação [7].

O estudo de técnicas que diminuam a quantidade de canais analisados em um sistema de identificação biométrica, pode impulsionar pesquisas para o desenvolvimento de aplicações reais. Uma revisão sistemática da literatura sobre seleção de canais para reconhecimento biométrico baseada em sinais de EEG foi realizada. Com base nesta revisão, foi evidenciada uma escassez de trabalhos relacionados à seleção de canais de EEG para fins biométricos, apresentando uma lacuna sobre a pesquisa de diferentes técnicas ainda inexploradas na área, tanto para a seleção de canais e extração de características quanto para o reconhecimento biométrico propriamente dito.

Além disso, o desenvolvimento de diferentes abordagens para a seleção de canais e extração de características dos

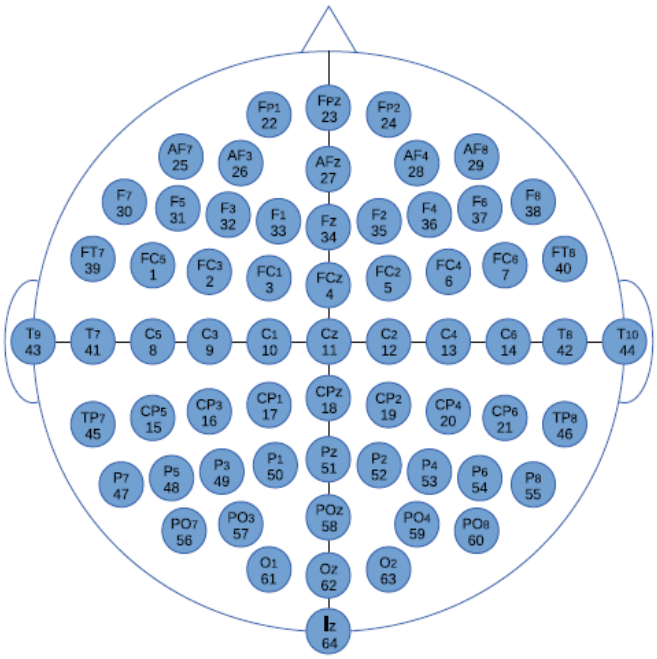


Figura 1: Sistema Internacional 10-10 padrão para posicionamento de sensores de EEG. [8]

sinais de EEG pode contribuir não apenas para a biometria. Essas técnicas podem ser adaptadas à outras áreas nas quais a otimização de canais e análise dos sinais de EEG seja necessária.

3. PROPOSTA DE SOLUÇÃO

Este projeto de pesquisa visa otimizar o processo de identificação biométrica baseada em EEG explorando diferentes estratégias para extração de características e seleção de canais mais relevantes visando reconhecimento biométrico.

Para o desenvolvimento do sistema de identificação biométrica baseada em EEG, serão consideradas quatro tarefas executadas na seguinte ordem: o pré-processamento dos sinais de EEG, a extração de características dos sinais já pré-precessados, a seleção de canais por meio de algoritmos de aprendizado de máquina e a classificação biométrica.

A partir do levantamento bibliográfico, o modelo AR se mostrou um dos mais utilizados. Segundo [1] o modelo autoregressivo (AR) é capaz de prover uma resolução superior para pequenos segmentos de dados, além de filtrar ruído branco. Por outro lado, a identificação de padrões periódicos pode revelar observações importantes sobre o comportamento e tendências futuras do problema representado pela série temporal e assim poderá conduzir a uma tomada de decisão eficaz. Essencialmente, a tarefa de simbolização consiste em segmentar o sinal em um conjunto de intervalos, os quais são particionados em classes e um símbolo é associado a cada classe. Esses símbolos são usados como entrada para o classificador. Em função disso, para a extração de características os estudos serão concentrados em dois algoritmos de análise de séries temporais, modelo autoregressivo (AR) e modelo de representação simbólica.

Três técnicas serão estudadas para a seleção de canais, duas técnicas já exploradas na literatura que servirão de

base de comparação (algoritmo genético e busca gulosa baseada em rankeamento) e a abordagem proposta (biclustering). No algoritmo genético, o cromossomo será representado pelos canais de EEG e para função de fitness será utilizado o classificador baseado em distância. A busca gulosa baseada em rankeamento utilizará técnicas heurísticas para seleção e teste dos melhores conjuntos de canais para identificação biométrica. Essas duas técnicas serão comparadas com a abordagem proposta baseada em biclustering.

A revisão sistemática realizada também permitiu a identificação das duas técnicas mais presentes na literatura para a classificação biométrica, Máquinas de Vetores de Suporte (SVM) e Redes Neurais Artificiais (RNA). Essas técnicas foram selecionadas pela alta capacidade de classificação ao tratar dados não linearmente separáveis. As Máquinas de Vetores de Suporte utilizam o truque de kernel para mapear o vetor de entrada para um espaço de características de alta dimensão. Já nas Redes Neurais Artificiais a separação dos dados é realizada por meio da combinação de funções sigmoides.

4. PROJETO DE AVALIAÇÃO DA SOLUÇÃO

A implementação das técnicas já utilizadas na literatura servirão para comparação com as técnicas propostas a fim de avaliar e validar a pesquisa. Primeiramente será avaliada a tarefa de seleção de características. As características extraídas pelo modelo AR e o modelo de representação simbólica serão avaliadas usando os classificadores SVM e RNA. Os resultados referentes a cada classificador serão comparados, em termos de taxa de reconhecimento, com base nas características extraídas de cada modelo.

Com as características selecionadas, a tarefa de seleção de canais será realizada. As abordagens com algoritmo genético, busca gulosa baseada em rankeamento e biclustering serão aplicadas e testadas com ambos os classificadores. Essas técnicas serão analisadas individualmente para identificação de valores ótimos, isto é, o menor número de canais que a técnica consegue selecionar mantendo uma boa acurácia na identificação biometrica. A partir desses resultados, será produzido um comparativo dessas três técnicas com os respectivos valores ótimos.

Será utilizada a base de dados pública disponível na internet chamada EEG Motor Movement/Imagery Dataset da Physionet [9, 6], o que facilitará a comparação e reprodução dos experimentos. Em todas as etapas técnicas estatísticas de significância serão utilizadas para avaliar a confiabilidade dos resultados encontrados.

5. ATIVIDADES JÁ REALIZADAS

A partir de uma busca exploratória na literatura sobre seleção de canais em biometria baseada em sinais de EEG, foi identificado uma ausência de trabalhos na área. Trabalhos secundários em biometria baseada em EEG não apresentam o problema de seleção de canais[3] ou apenas o citam como lacuna para futuras pesquisas [2].

Com base esses resultados, uma revisão sistemática sobre seleção de canais em biometria baseada em EEG foi desenvolvida. O objetivo da revisão foi analisar sistematicamente o estado da arte sobre o problema de seleção dos melhores canais de EEG para reconhecimento biométrico. Foi analisado artigos que explorem o tema, identificando técnicas utilizadas na literatura para seleção de canais, bem como as bases de dados usadas e as abordagens já exploradas para a tarefa de classificação biométrica e extração de características.

A busca na literatura retornou 70 trabalhos distintos que foram reduzidos a 17 artigos aceitos, os quais realizaram seleção de canais de EEG para biometria. Os dados foram extraídos com base em quatro aspectos de interesse: bases de dados utilizadas, métodos para extração de características, métodos para seleção de canais e métodos para classificação biométrica.

A análise em relação às bases de dados mostrou 3 bases de dados públicas disponíveis na internet com potencial para uso neste projeto. A base de dados EEG Motor Movement/Imagery Dataset da Physionet se mostrou a mais robusta e foi selecionada para o projeto. Essa base possui dados de 64 canais coletados a partir de 109 indivíduos diferentes com uso de quatro protocolos: estado de repouso com olhos fechados, estado de repouso com olhos abertos, tarefa motora e tarefa motora imaginária.

Os artigos da revisão sistemática apresentaram 13 técnicas distintas para extração de características. O modelo autoregressivo foi o mais utilizado, aparecendo em 5 artigos. Por esse motivo foi escolhido como base de comparação para a técnica proposta nesse projeto. Quanto à seleção de canais, pode-se identificar uma predominância de técnicas de busca heurística e meta-heurística para combinação de diferentes configurações de canais.

Já para a tarefa de classificação biométrica, pode-se observar uma tendência a identificação biométrica, 14 artigos trabalham com essa abordagem em comparação a 3 trabalhos de autenticação. Dentre os métodos para classificação biométrica a utilização de SVM se mostrou predominante, com 8 artigos utilizando esta técnica para classificação. A segunda técnica mais utilizada foi Redes Neurais Artificiais, exploradas em 7 artigos.

6. CONCLUSÃO

Este trabalho está inserido no contexto de aperfeiçoamento de sistemas de identificação biométricos. Esperar-se que a contribuição desse trabalho ao aplicar diferentes técnicas para otimizar de canais visando identificação biométrica, possa ajudar na evolução da pesquisa para aplicações práticas da biometria baseada em EEG.

Espera-se também que os resultados de classificação biométrica dos métodos que serão desenvolvidos, apresentem uma acurácia melhor ou similar aos métodos de artigos prévios que utilizem todos os canais, além de identificar os canais mais relevantes para o problema. Com isso, essas técnicas poderiam influenciar o desenvolvimento de tecnologias para captura de sinais de EEG mais adequadas para sistemas práticos de biometria.

Os métodos desenvolvidos nesse trabalho têm como finalidade a otimização da classificação biométrica, no entanto, são adaptáveis para aplicações em outras áreas na qual a seleção de canais representativos do sinal de EEG seja uma questão importante. As técnicas desenvolvidas neste projeto podem ser aplicadas a problemas relacionados à interface cérebro-computador (brain–computer interface - BCI) ou até na medicina em diagnósticos de doenças neurológicas.

7. REFERÊNCIAS

[1] M. K. Abdullah, K. S. Subari, J. L. C. Loong, and N. N. Ahmad. Analysis of effective channel placement

for an eeg-based biometric system. In *2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 303–306, Nov 2010.

[2] M. Abo-Zahhad, S. M. Ahmed, and S. N. Abbas. State-of-the-art methods and future perspectives for personal recognition based on electroencephalogram signals. *IET Biometrics*, 4(3):179–190, 2015.

[3] P. Campisi and D. L. Rocca. Brain waves for automatic biometric-based user recognition. *IEEE Transactions on Information Forensics and Security*, 9(5):782–800, May 2014.

[4] P. Cserti, B. Végső, G. Kozmann, Z. Nagy, F. D. V. Fallani, and F. Babiloni. Methods to highlight consistency in repeated {EEG} recordings. {*IFAC*} *Proceedings Volumes*, 45(18):23 – 27, 2012. 8th {IFAC} Symposium on Biological and Medical Systems.

[5] M. Fraschini, A. Hillebrand, M. Demuru, L. Didaci, and G. L. Marcialis. An eeg-based biometric system using eigenvector centrality in resting state brain networks. *IEEE Signal Processing Letters*, 22(6):666–670, June 2015.

[6] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000.

[7] K. V. R. Ravi and R. Palaniappan. A minimal channel set for individual identification with eeg biometric using genetic algorithm. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, volume 2, pages 328–332, Dec 2007.

[8] D. Rodrigues, G. F. Silva, J. P. Papa, A. N. Marana, and X.-S. Yang. Eeg-based person identification through binary flower pollination algorithm. *Expert Systems with Applications*, 62:81 – 90, 2016.

[9] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

[10] S.-K. Yeom, H.-I. Suk, and S.-W. Lee. Person authentication from neural activity of face-specific visual self-representation. *Pattern Recognition*, 46(4):1159 – 1169, 2013.

Reconhecimento de padrões aplicado a análise de expressões faciais gramaticais da língua brasileira de sinais: uma abordagem usando mistura de especialistas

Pattern recognition applied to grammatical facial expressions analysis in brazilian sign language: a mixtures of experts approach

Maria Eduarda de A. Cardoso
Universidade de São Paulo
Av. Arlindo Bettio, 1000
03828-000 São Paulo, Brasil
dudaa@usp.br

Sarajane M. Peres
Universidade de São Paulo
Av. Arlindo Bettio, 1000
03828-000 São Paulo, Brasil
sarajane@usp.br

RESUMO

O reconhecimento das expressões faciais tem sido um grande atrativo para pesquisadores de diferentes áreas, pois tem um grande potencial para desenvolvimento de aplicações. Reconhecer automaticamente essas expressões tem se tornado um objetivo importante na área de análise do comportamento humano. Sobretudo em estudo das línguas de sinais, a análise das expressões faciais representa um fator importante, por tratar-se de uma língua de modalidade visual-espacial que não tem o suporte sonoro, e, portanto, necessita de algum elemento gestual para expressar informações de prosódia e ajudar no suporte ao desenvolvimento da sua estrutura gramatical. Nesse contexto, as expressões faciais são chamadas de expressões faciais gramaticais. Entre as linhas de estudos que exploram essa temática, está a análise automática da língua de sinais. Assim, essa pesquisa propõe desenvolver uma arquitetura capaz de realizar a identificação das expressões em uma sentença da língua brasileira de sinais (Libras), segmentando tal sentença de acordo com cada tipo diferente de expressão usada em sua construção.

Palavras-Chave

Expressão facial gramatical, reconhecimento de padrões, língua brasileira de sinais, língua de sinais, mistura de especialistas.

ABSTRACT

Researchers from different fields have become interested in recognition of facial expressions, since this area has a huge potential to support the development of applications. Recognizing facial expressions by automatic means has become a goal mainly in the human behavior analysis field. Specially in sign language studies, analysis of facial expressions represents an important issue, since such languages assume a visual-spatial form and does not have sound support. It requires some gestural element in order to express information in prosody and also to help in the development of its grammatical structure. In this context, the facial expressions are called grammatical facial expressions. Among the research lines that address this theme, there is one that aims at the implementation of automatic analyzes of sign languages. In our research, we propose to develop an architecture capable of accomplishing the identification of grammatical facial expressions within a sentence of Brazilian Sign Language (Libras), carrying out the segmentation of a sentence according to each facial expression used in its construction.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•Information systems → Data mining; •Computing methodologies → Supervised learning by classification;

Keywords

Grammatical facial expressions, pattern recognition, brazilian sign language, sign language, mixtures of experts.

1. INTRODUÇÃO

Uma das maneiras mais representativas pela qual o ser humano demonstra seus sentimentos é por meio de expressões faciais (EF). Expressões faciais emitem emoções e assim, por meio da análise das expressões é possível reconhecer emoções e dotar agentes de software da capacidade de usar essa informação na melhoria da interação humano-computador. Segundo estudos da Psicologia, toda manifestação de uma expressão facial resulta da ocorrência de uma emoção, mesmo no caso de uma expressão neutra. Há conjuntos de emoções possíveis, gerado a partir das relações e reações emocionais, e são suficientes para que se possa compreender as relações entre os seres humanos [15]. Esse conjunto é composto por seis emoções: felicidade, surpresa, raiva, desgosto, medo e tristeza. Estudos recentes defendem que essas seis

emoções podem ser resumidas em quatro: felicidade, tristeza, medo/surpresa e desgosto/raiva [6].

No contexto das línguas de sinais (LSs), as EF assumem um papel extra à expressividade da emoção de um indivíduo, pois elas são usadas na formação da estrutura gramatical da língua. Um dos primeiros estudos a formalizar a estrutura da LSs foi realizado por [13], em 1960 e nele as EF já foram inseridas com um dos elementos constituintes da língua. Quando as expressões são usadas na composição da estrutura de uma língua de sinais elas são chamadas de expressões faciais gramaticais (EFG). As EFGs estão presentes nos níveis morfológicos da língua, no qual têm atribuição de adjetivação, e no nível sintático, que é responsável por determinar o sentido das frases. As EFGs no contexto da língua brasileira de sinais (Libras) são: interrogativa, negativa, afirmativa, condicional, relativa, tópicos e foco. Alguns estudos têm mostrado melhorias no desempenho de modelos de reconhecimento automático de sinais em LSs quando implementam abordagens multimodais de análise e incluem o reconhecimento das EFGs. São exemplos desses estudos: [11] e [8], que trabalham na interpretação da Língua de Sinais Americana, e os estudos de [16] e [1] que versam sobre o uso de Máquinas de Vetores Suporte para classificação de EFG.

2. APRESENTAÇÃO DO PROBLEMA

O objetivo desta pesquisa é construir um classificador capaz de reconhecer uma EFG em uma sentença expressa em Libras. Trata-se de um projeto que estende um projeto já realizado [5] dentro do grupo de pesquisa associado, que tratou o mesmo problema porém sob uma perspectiva de um problema de classificação binário. O presente projeto suprirá a lacuna deixada pelo projeto anterior, qual seja, implementar o reconhecimento das EFGs sob uma perspectiva de um problema de classificação multiclasse. O problema de reconhecimento de padrões será estudado sob dois diferentes aspectos, atemporal (um quadro de vídeo) e temporal (conjunto de quadros de vídeo). Além disso, uma organização de dados em janelas deverá ser realizado para que seja possível caracterizar a informação sobre a movimentação dos elementos da face no tempo. A tabela 1, mostra um esquema que explica a construção de janelas de tamanhos diferentes com conjuntos de quadros de vídeos, vale a pena levar em conta as informações sobre quadros anteriores e posteriores na análise de cada quadro, para obter informações temporais.

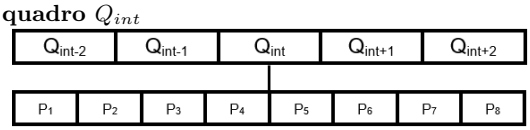
Tabela 1: Janelas de tamanhos distintos

Tam.	Janela 1	...	Janela n
1	$quadro_1$	...	$quadro_n$
2	$quadro_1$;$quadro_2$	...	$quadro_{n-1}$;$quadro_n$

Na figura 1 é ilustrada a estrutura da janela, incluindo as características de um quadro. Nesta figura estão os oitos pontos selecionados dos cem pontos extraídos da face humana por meio do sensor Kinect. Os pontos P_1, P_2, P_3 e P_4 são referentes a sobrancelha, os pontos P_5, P_6, P_7 e P_8 pertencem a boca. É importante ressaltar que o conjunto de pontos obtidos está em conformidade com estudos da extração de EF como [3] e [14].

A definição desse problema é uma extensão do problema

Figura 1: Exemplo de janela com 8 pontos centrado no quadro Q_{int}



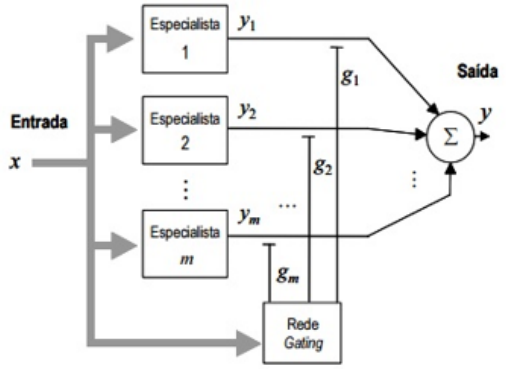
que foi apresentado por [5]. Neste projeto, uma expressão facial EFG_i, pertencente ao conjunto finito de expressões faciais EFG_1, EFG_2, ..., EFG_n, é descrita por um conjunto de pontos P = p_1, p_2, ..., p_n extraídos da face humana e dispostos no espaço tridimensional. A figura 2 mostra uma face neutra, uma face executando uma EFG usada na Libras e a plotagem do conjunto de pontos que descreve uma face.

Figura 2: Face neutra, face com uma execução de uma EFG, e os respectivos pontos (x,y) extraídos da face [5]

A fim de definir o problema básico a ser estudado durante a realização deste projeto, considere um vídeo como sendo uma sequência de quadros S = q_1, q_2, ..., q_t de tamanho t. Uma representação vetorial dos quadros desse vídeo é usada como entrada para um modelo de classificação multiclasse sob uma arquitetura de mistura de especialistas (ME) composta por multilayer perceptrons (MLP), cujo objetivo é classificar cada quadro como sendo referente à execução de uma EFG específica durante a execução de uma fala em Libras. Misturas de especialistas é uma arquitetura modular para aprendizado supervisionado e foi devidamente formalizada por [7]. Trata-se de um método que tenta solucionar problemas de classificação ou regressão com base em uma estratégia dividir-e-conquistar – dividindo o problema entre vários especialistas (no caso deste projeto, os especialistas são modelos classificadores construídos com MLP). Na ME o espaço de entrada é automaticamente dividido em regiões, sendo que para cada região existe um único ou um subconjunto de especialistas mais indicados a agir. A figura 3 é ilustrada a estrutura de uma arquitetura de especialistas. As MLPs são redes neurais que surgiram a partir da criação de um modelo de neurônio artificial chamado Perceptron [12]. MLPs são ótimas detectoras de características, devido suas camadas ocultas que são formadas por neurônios do tipo Perceptron interconectados, responsáveis por realizarem localmente, e de forma eficiente, a discretização do erro envolvido na tarefa de aprendizado. Existem outros aprendizados indutivos, como as redes recorrentes que foram utilizadas em [2]. Nesse estudo para classificar as expressões faciais gramaticais, foi utilizada uma rede 25 neurônios na camada de entrada, 20 neurônios na camada oculta e a camada de saída tantos neurônios quanto as classes possíveis correspondentes às expressões faciais utilizadas no trabalho.

A hipótese desta pesquisa baseia-se no fato que EFGs podem ser automaticamente localizadas e identificadas dentro de uma sentença da Libras com o apoio da mistura de especialistas. Desde que cada especialista é definido automaticamente de acordo com o problema sob resolução, acredita-se que a ME seja capaz de direcionar um ou mais especialistas para reconhecer uma EFG.

Figura 3: Estrutura típica de uma arquitetura de mistura de especialistas [4]



3. PROPOSTA DE SOLUÇÃO

Este projeto está organizado como uma pesquisa do tipo experimental, que conterá pesquisa bibliográfica exploratória, revisão sistemática, revisão de conjunto de dados de referência, implementação e teste de algoritmos e avaliação de resultados referentes à construção de modelos classificadores para localização das EFGs na sentença em Libras. Será realizada a revisão do conjunto de dados *Grammatical Facial Expression*[1], incluindo a realização de uma nova rotulação das sentenças presentes no conjunto. Este conjunto de dados já foi utilizado nos experimentos descritos em [5]. Ele é composto por dezoito arquivos de vídeos gravados utilizando o sensor Microsoft Kinect, com apoio de funções disponibilizadas na *Face Tracking SDK*[2]. Em cada vídeo, um usuário executa (cinco vezes), em frente ao sensor, cinco frases em Libras que exigem o uso de uma expressão facial gramatical. Usando o Microsoft Kinect, os autores desse conjunto de dados obtiveram: (a) uma imagem de cada quadro, identificado por um marcador de tempo; (b) um arquivo de texto contendo coordenadas (x, y, z) de pontos de olhos, nariz, sobrancelhas, contorno do rosto e da boca. O procedimento realizado para construção desse conjunto de dados deverá ser repetido neste trabalho para que novas sentenças da Libras, que usem mais de uma expressão facial, sejam incorporadas ao conjunto. Haverá a execução de uma nova coleta de dados para a experimentação, um protocolo será estabelecido e formalizado em projeto para ser submetido a um comitê de ética em pesquisa. A coleta de dados envolverá a participação de seres humanos, havendo necessidade de captação da imagem das pessoas e de submetê-las a um processo com risco de causar algum tipo de estresse. Os quadros de vídeos não estarão disponíveis publicamente mas estarão disponíveis para uso dentro do escopo do grupo de pesquisa. A proponente desta pesquisa possui conhecimentos básicos em Libras e deverá realizar o trabalho de revisão do conjunto junto de especialistas que trabalham no departamento de linguística da FFLCH/USP. A construção dos classificadores se dará em dois momentos. Num primeiro momento a técnica MLP será aplicada no problema (completo e em versões simplificadas do problema) fora da arquitetura de ME, para que seja possível melhorar o conhecimento sobre a complexidade da resolução de um problema multiclasse nessa área. Em seguida serão implementadas as arquiteturas de misturas de especialistas fazendo uso também das técnicas citadas.

[1]https://archive.ics.uci.edu/ml/datasets/Grammatical+Facial+Expressions

[2]https://msdn.microsoft.com/en-us/library/jj130970.aspx

4. AVALIAÇÃO DA SOLUÇÃO

A avaliação dos classificadores será realizada via medidas tradicionalmente usadas na área de reconhecimento de padrões, como a métrica de matriz de confusão, que oferece uma medida efetiva do modelo de classificação ao mostrar o número de classificações corretas e as classificações preditas para cada classe em um determinado conjunto [10]. Também será realizada a avaliação de acordo com a visão dos especialistas. De maneira mais detalhada, a forma de aferir o erro de classificação pode ser por meio do erro total cometido pelo classificador em termos de número de quadros classificados erroneamente e a porcentagem que ela representa do total de quadros apresentados no teste. Essa avaliação deverá ainda ser enriquecida com a análise de erros de borda e erros de segmentação [9]. Os erros de borda são definidos como os erros de classificação que ocorrem dentro da faixa de transição. O erro acontece entre a ocorrência da EFG sob análise e a não ocorrência, que é a "não expressão". No caso das expressões faciais gramaticais é interessante avaliar o desempenho do classificador sob o ponto de vista de um especialista que usará o resultado produzido pelo algoritmo. Assim, a avaliação quantitativa de desempenho dos modelos classificadores deverá ser combinada com avaliações qualitativas por meio de um protocolo de avaliação que considere também a visão de um educador em língua de sinais interessado no estudo da execução da EF ou da interpretação da Libras.

5. ATIVIDADES JÁ REALIZADAS

Como parte do desenvolvimento do projeto, foram realizados estudos exploratórios referentes à área de Língua de Sinais, MLP e ME. Uma revisão sistemática sobre reconhecimento de padrões em EF afetivas e em EFG foi realizada. Na condução foram encontrados muitos artigos no qual passaram por leitura, e por critérios e análises, onde haviam questões de pesquisas na qual os artigos tinham que abordar essas respostas, para serem inclusos na revisão. Nos estudos realizados é perceptível que a área de reconhecimento de expressões faciais vem sendo exploradas, de forma mais frequente, a partir do uso de técnicas como Random Forest, Máquinas de Vetores Suporte e MLP. Conjuntos de dados como Jaffe, Cohn-Konade e MUG, foram utilizadas em vários trabalhos. A técnica de matriz de confusão para avaliação de desempenho é utilizada pela maioria dos trabalhos. O conjunto de dados já existente é original rotulado de forma binária. A fim de suportar os primeiros experimentos deste projeto, antes da realização da nova coleta de dados, o conjunto original foi reorganizado porém usando múltiplas

classes. Para isso, as sentenças na organização original foram combinadas na nova organização gerando um único conjunto de dados multiclasse, de forma que as classes tenham rótulos diferentes. No trabalho anterior classificado binário, quando ocorria EFG na frase era rotulado 1 e 0 quando não ocorria EFG. No presente trabalho os rótulos multiclasse, seguem a seguinte rotulação: quando ocorre EFG afirmativa é rotulado 1, quando ocorre EFG negativa é 2, quando ocorre EFG interrogativa é 3 e assim sucessivamente. A rotulação do trabalho anterior, e a nova rotulação para o presente trabalho não são livres de viés pois são rotuladas por seres humanos especialistas em língua de sinais, e assim existindo uma subjetividade, pois cada especialista rótula seus dados com sua percepção. Esse conjunto de dados já está preparado e pré-processado e neste momento está sendo usado para testes da aplicação de MLPs (ainda fora da mistura de especialistas).

6. CONCLUSÃO

O objetivo deste projeto é construir uma arquitetura capaz de realizar a identificação das expressões faciais gramaticais em uma sentença em Libras e segmentar essa sentença de acordo com cada expressão. A apresentação desta arquitetura poderá suportar estudos na área de Linguística que se baseiam em expressões faciais da Libras para elucidar questões inerentes ao sistema linguístico, como um sistema multimodal onde EFs representam um elemento provido de significado, de informação e de capacidade de comunicação. O método de análise das EF estará pautado em técnicas de aprendizado de máquina supervisionado e requererá um estudo sobre modelagem de classificadores e mistura de especialistas. Tais realizações devem contribuir para a área de análise automática de EF. Uma vez que representações e estratégias de análise propostas para um fim específicos podem transitar entre diferentes domínios de aplicação, os progressos deste projeto se configuram como uma contribuição para a área de reconhecimento de padrões.

7. REFERÊNCIAS

[1] I. Ari, A. Uyar, and L. Akarun. Facial feature tracking and expression recognition for sign language. In *Int. Symp. on Comp. and Inf. Sci.*, pages 1–6. IEEE, 2008.

[2] G. Caridakis, S. Asteriadis, and K. Karpouzis. Non-manual cues in automatic sign language recognition. *Personal and ubiquitous computing*, 18(1):37–46, 2014.

[3] C.-Y. Chang and Y.-C. Huang. Personalized facial expression recognition in indoor environments. In *Int. J. Conf. on Neural Networks*, pages 1–8. IEEE, 2010.

[4] C. A. de Moraes Lima. *Comitê de Máquinas: uma abordagem unificada empregando máquinas de vetores-suporte*. PhD thesis, Universidade Estadual de Campinas, 2004.

[5] F. d. A. Freitas. Reconhecimento automático de expressões faciais gramaticais na língua brasileira de sinais. Master's thesis, Univ. de São Paulo, 2015.

[6] R. E. Jack, O. G. Garrod, and P. G. Schyns. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time, 2014.

[7] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural comput.*, 3(1):79–87, 1991.

[8] H. Kacorri and M. Huenerfauth. Continuous profile models in asl syntactic facial expression synthesis. ACL, 2016.

[9] R. C. B. Madeo. Máquinas de vetores suporte e a análise de gestos: incorporando aspectos temporais, 2013.

[10] M. C. Monard and J. A. Baranauskas. Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, 1(1), 2003.

[11] T. D. Nguyen and S. Ranganath. Facial expressions in american sign language: Tracking and recognition. *Pat. Recog.*, 45(5):1877–1891, 2012.

[12] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psy. Rev.*, 65(6):386, 1958.

[13] W. C. Stokoe. Sign language structure: An outline of the visual communication systems of the american deaf. *J. of Deaf Studies and Deaf Education*, 10(1):3–37, 2005.

[14] H. Wang, H. Huang, Y. Hu, M. Anderson, P. Rollins, and F. Makedon. Emotion detection via discriminative kernel method. In *Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments*, page 7. ACM, 2010.

[15] C. Whissell, M. Fournier, R. Pelland, D. Weir, and K. Makarec. A dictionary of affect in language: Iv. reliability, validity, and applications. *Perceptual and Motor Skills*, 62(3):875–888, 1986.

[16] H.-D. Yang and S.-W. Lee. Combination of manual and non-manual features for sign language recognition based on conditional random field and active appearance model. In *Int. Conf. on Machine Learn. and Cyber.*, volume 4, pages 1726–1731. IEEE, 2011.

Coagrupamento de dados para melhoria da serendipidade em sistemas de recomendação baseados em conteúdo

Alternative Title: Co-clustering for serendipity improvement in content-based recommender systems

Andrei Martins Silva
Universidade de São Paulo
03828-000, São Paulo - SP
andreimartins@usp.br

Sarajane Marques Peres
Universidade de São Paulo
03828-000, São Paulo - SP
sarajane@usp.br

RESUMO

Sistemas de recomendação baseados em conteúdo têm sido amplamente utilizados para recomendação de itens em vários domínios de aplicação como entretenimento, comércio eletrônico e notícias. No entanto, uma de suas principais desvantagens é a falta de serendipidade das recomendações, problema que expõe seus usuários a itens excessivamente similares ao seu perfil. O coagrupamento de dados é uma técnica capaz de encontrar relações de similaridade parcial entre itens, as quais podem ser exploradas para oferecimento de recomendações serendipitosas. Este trabalho propõe o uso de coagrupamento de dados para construção de um sistema de recomendação baseado em contéudo que apresente boa característica de serendipidade. Experimentos preliminares conduzidos sobre um conjunto de dados de notícias indicam o potencial da solução proposta.

Palavras-Chave

Sistemas de recomendação baseados em conteúdo, serendipidade, coagrupamento de dados

ABSTRACT

Content-based recommender systems have been widely used for item recommendations in several application domains such as entertainment, e-commerce and news. However, one of its major drawbacks is the lack of serendipity in recommendations, problem which exposes its users to excessively similar items to their profile. Co-clustering is a method capable of finding partial similarity among items which can be exploited to offer serendipitous recommendations. This work proposes the use of co-clustering to build a serendipitous content-based recommender system. Preliminary experiments conducted over a news dataset suggests the potential of the proposed solution.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

Keywords

Content-based recommender systems, serendipity, co-clustering

CCS Concepts

•Information systems → Recommender systems; •Computing methodologies → Non-negative matrix factorization; •Theory of computation → *Unsupervised learning and clustering;*

1. INTRODUÇÃO

Sistemas de recomendação auxiliam os usuários a lidar com o problema de sobrecarga de informações [17]. Tais sistemas fornecem recomendações de itens como filmes, músicas e notícias com base no perfil de cada usuário. O sistema de recomendação da Amazon [12], por exemplo, recomenda produtos que possam interessar a seus usuários com base na similaridade entre seu histórico de compras e novos itens de seu catálogo. Por sua vez, a Netflix sugere filmes que possam agradar ao usuário com base nos filmes que ele assistiu no passado [3].

A recomendação baseada em conteúdo é uma das mais bem sucedidas abordagens para construção de sistemas de recomendação. A ideia básica é construir um perfil de usuário a partir de palavras-chave contidas em seus itens prediletos. Então, por meio de uma medida de similaridade, como a similaridade de cossenos, o sistema compara esse perfil com itens do catálogo ainda não vistos pelo usuário. Os itens considerados mais similares ao perfil do usuário são recomendados [1].

Apesar de seu sucesso, sistemas que adotam essa abordagem tendem a sofrer do problema da falta de serendipidade, em que apenas itens muito similares ao perfil do usuário são recomendados [2]. Serendipidade é uma característica desejável em sistemas de recomendação, pois propicia aos usuários recomendações relevantes e inesperadas, as quais não seriam encontradas por ele de maneira autônoma [10].

O coagrupamento de dados é a estratégia proposta neste trabalho para solução do problema da serendipidade em sistemas de recomendação baseados em conteúdo, pois explora a *dualidade* [7] existente na relação entre itens e suas palavras-chave, sendo capaz de identificar, simultaneamente, *cogrupos* de itens e palavras-chave. Assim torna-se possível revelar relações do tipo: "item A é similar ao item B quando considerados apenas os termos 1, 2, 3, 38, 39, mas é dissi-

milar quando considerados os demais".

Nos últimos anos, pesquisadores têm se esforçado para construir sistemas de recomendação baseados em conteúdo que forneçam recomendações serendipitosas. Uma revisão sistemática cujos detalhes são apresentados na seção 5, não encontrou trabalhos que utilizassem o coagrupamento de dados para solução do problema da serendipidade. Ao invés disso, o coagrupamento de dados tem sido usado para resolver outros problemas de recomendação como acurácia [4] e *cold-start* [15].

Entre as propostas identificadas para solucionar o problema da serendipidade, destacam-se duas vertentes: aumento da diversidade e enriquecimento de perfis. A introdução de aleatoriedade no processo de recomendação foi indicada como abordagem para aumentar a diversidade das recomendações, e consequentemente, para a melhoria da serendipidade [2]. A fragilidade desta abordagem é que, ao recomendar itens mais surpreendentes, aumenta-se o risco de recomendar itens irrelevantes, pois geralmente os itens surpreendentes estão mais distantes do perfil do usuário [10, 20]. A segunda vertente, busca enriquecer os perfis de itens e usuários com informações externas oriundas de enciclopédias, ontologias ou *folksonomias*. Nesta linha, em [6], foram utilizados conhecimentos enciclopédico (Wikipédia) e linguístico (*WordNet*) combinados a um algoritmo de caminhada aleatória para solucionar o problema da serendipidade. Similarmente, em [19], são utilizadas *folksonomias* para enriquecimento do conteúdo dos itens visando recomendações serendipitosas. Além das *folksonomias*, metadados como gêneros (de filmes) e palavras-chave foram propostos em [14] para enriquecimento do conteúdo dos itens em busca da criação de um modelo que supere a superespecialização nas recomendações. Apesar dos resultados serem promissores, o enriquecimento dos perfis, seja por meio de ontologias ou *folksonomias*, representa um fator extra de complexidade ao modelo.

Seja por meio da inserção de aleatoriedade à recomendação ou pelo enriquecimento de perfis, alcançar a serendipidade em sistemas de recomendação baseados em conteúdo permanece uma questão em aberto [6], e, portanto, é o problema central tratado nesta pesquisa. Ao contrário de abordagens que inserem aleatoriedade ao processo de recomendação, o coagrupamento de dados melhora a surpresa das recomendações por meio da identificação de similaridades parciais, o que, ao mesmo tempo, lhe permite também manter a relevância dos itens recomendados. Estes dois fatores - surpresa e relevância - levam à recomendações serendipitosas. Além disso, o modelo é mais simples, quando comparado a modelos que fazem uso de conhecimento externo como os apresentados em [14, 19, 6], pois opera somente sobre as descrições textuais dos itens.

2. PROBLEMA

Considere $H \in \Re^{p \times n}$ a matriz de interações de p usuários de um sistema de recomendação com os n itens de seu catálogo. O elemento $\mathrm{w}(u_i,c_j) \in H$ assume valores binários ou numéricos que indicam a relação entre o usuário u_i e item c_j. Além disso, considere $C = \{c_1, \ldots, c_n\}$ o conjunto de itens do catálogo de recomendação. Então, um sistema de recomendação baseado em conteúdo pode ser visto como uma função $\phi : H \times C \rightarrow L$ em que $L = \{l_1, l_2, l_3, \ldots, l_q\} \subset C$ representa o conjunto de listas de recomendação fornecidas a seus usuários. Uma lista de recomendação l_k, contendo N itens é conhecida como lista *Top-N*, pois apresenta os N itens supostamente mais relevantes para o usuário. Seja $S(l_k) \rightarrow \nu_k$ uma medida apropriada de serendipidade de uma lista de recomendação k, o problema enfrentado pelo sistema de recomendação em questão é encontrar o conjunto de listas de recomendação L que maximize a medida de serendipidade $S(L)$.

3. SOLUÇÃO PROPOSTA

Cada item $c \in C$ apresentado na seção 2 pode ser representado por um vetor m-dimensional de palavras-chave. Por exemplo, uma notícia de jornal poderia ser representada pelo vetor $c_1 = [\text{temporal} = 1, \text{derruba} = 1, \ldots, \text{esplanada} = 0, \ldots]$. Seguindo esta formalização, o conjunto C pode ser representado como uma matriz, em que cada linha representa um termo e cada coluna um item. Então, algoritmos de agrupamento poderiam ser aplicados para encontrar grupos de notícias similares e, a partir dos grupos, formar as listas de recomendação. Analogamente, o coagrupamento de dados pode ser aplicado de modo a encontrar cogrupos que definem itens parcialmente similares.

Na literatura, algoritmos de fatoração de matrizes têm sido utilizados com sucesso para coagrupamento de dados em altas dimensões, como são os itens tratados neste trabalho [18, 13]. A fatoração de matrizes consiste em decompor a matriz C em fatores que, quando multiplicados, são capazes de aproximar a matriz original: $C \approx UV^T$. Quando a decomposição é realizada utilizando três fatores - $C \approx USV^T$ - o processo é chamado trifatoração. Neste trabalho, a decomposição em dois fatores é implementada pelo algoritmo *Non-negative Matrix Factorization (NMF)* [18] e a decomposição em três fatores pelo algoritmo *Non-negative Block Value Decomposition (NBVD)* [13].

NMF e NBVD são algoritmos guiados pelo processo de *minimização* de funções-objetivo que captam a distância entre a aproximação e a matriz original $X \in \Re^{m \times n}$.

$$\min_{U,V} \frac{1}{2}\|X - UV^T\| \qquad \text{(NMF)}$$

$$\min_{U,S,V} \|X - USV^T\|^2 \qquad \text{(NBVD)}$$

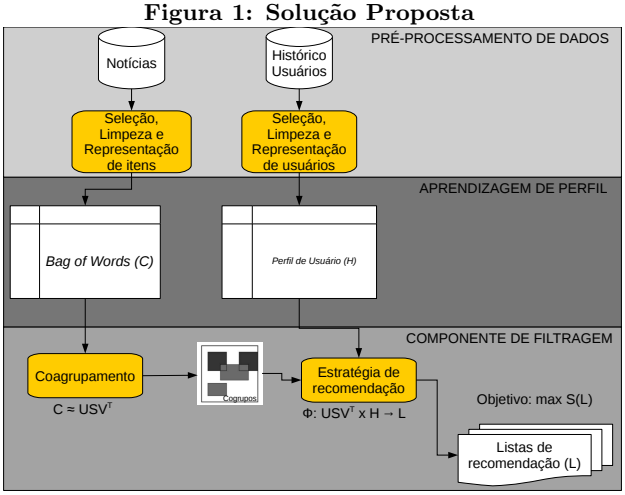
em que $\forall ij : U_{ij}, V_{ij}, S_{ij} \geq 0$, $U \in \Re^{m \times k}$, $S \in \Re^{k \times l}$, $V \in \Re^{n \times l}$ e $\|\cdot\|$ denota a norma de Frobenius.[1]

No caso do coagrupamento de documentos, a matriz X possui m termos e n documentos. Nesta configuração, a matriz U representa a pertinência do i-ésimo termo ao j-ésimo grupo. Por sua vez, a matriz V representa a pertinência do i-ésimo documento ao j-ésimo grupo. S pode ser entendida como uma versão compacta da matriz original X [13].

A figura 1 ilustra a solução proposta. Os conjuntos de notıcias e usuários passam por um processo de seleção e limpeza dando origem, respectivamente, às matrizes C e H. Em seguida, um algoritmo de fatoração de matrizes nãonegativas é aplicado à C de modo a obter cogrupos de notícias parcialmente similares - $C \rightarrow U, V$ ou $C \rightarrow U, S, V$. Então, os cogrupos encontrados são combinados com o perfil dos usuários para geração de listas de recomendação - $U, V, H \rightarrow L$ ou $U, S, V, H \rightarrow L$. Os resultados experimentais descritos neste artigo referem-se à aplicação de algoritmos de fatoração de matrizes não-negativas à matriz C. A

[1]Na formulação do NMF, $k = l$ obrigatoriamente.

Figura 1: Solução Proposta



estratégia de recomendação complementa a solução e será implementada e avaliada no futuro.

4. AVALIAÇÃO

A estratégia de recomendação para alcançar a serendipidade proposta nesta pesquisa será validada sobre bases de dados cujos itens estejam descritos por seu conteúdo textual. São previstas duas formas de avaliação do desempenho da solução:

Intrínseca Essa avaliação buscará avaliar a capacidade da solução proposta de fornecer recomendações serendipitosas de acordo com métricas apropriadas encontradas na literatura como a SRDP [8]. Em [5], a solução proposta apresenta valores entre 0,4 e 0,7 na métrica SRDP. Outra medida de serendipidade - Serendipity@N - é proposta em [6]. Os autores reportam valores entre 0,02 e 0,16 alcançados por diferentes algoritmos avaliados[2]. Tais avaliações constituem-se como importantes bases de comparação, mesmo com a ressalva de que as condições de experimentação não serão exatamente iguais. A validação interna dos cogrupos será aferida por índices como o de Dunn e Davies-Bouldin [9]. Análises estatísticas como testes de hipóteses paramétricos e/ou não-paramétricos, bem como estabelecimento de intervalos de confiança serão aplicados para suportar a avaliação.

Extrínseca Serendipidade é um conceito difícil de medir [10], pois envolve aspectos que nem sempre podem ser capturados de maneira objetiva como a surpresa causada por uma recomendação. Por este motivo, essas avaliações serão realizadas com base na interação de usuários e um protótipo computacional da solução disponibilizado na Internet[3], aplicação de questionários e análises qualitativa e quantitativa dos resultados.

[2]Ambas as métricas podem assumir valores entre 0 e 1. Quanto maior, melhor.

[3]Com as devidas permissões concedidas pelo Conselho de Ética da Universidade de São Paulo.

5. ATIVIDADES REALIZADAS

O projeto de pesquisa está estruturado em quatro macro-atividades descritas nesta seção, incluindo o avanço, em percentual, de cada atividade. Em seguida, resultados referentes a experimentos preliminares também são apresentados.

Revisão bibliográfica Estudos exploratórios (70%), revisão sistemática (RS) sobre serendipidade em sistemas de recomendação (66%). A RS tem como objetivo identificar o estado da arte na solução do problema da serendipidade em sistemas de recomendação. Seguindo os procedimentos descritos em [11], palavras-chave como *recommender systems, filtering, personalization, algorithms, serendipity, over-specialization* foram pesquisadas nas bases eletrônicas ACM, IEEE, Springer, Scopus e Web Of Science para identificação de trabalhos relevantes. Em seguida, critérios de inclusão e exclusão foram aplicados e os 64 artigos restantes analisados na íntegra.

Implementação Proposição da solução (50%), pré-processamento de dados (30%), implementação de algoritmos de coagrupamento (60%), implementação da estratégia de recomendação (0%).

Avaliação Experimentos preliminares (80%), avaliação intrínseca (0%), avaliação extrínseca (0%).

Divulgação Publicação de resultados em *workshops*, conferências e periódicos.

A fim de verificar o potencial do coagrupamento de dados para melhoria da serendipidade em sistemas de recomendação baseada em conteúdo um experimento preliminar foi conduzido sobre uma base de dados real.

A base de dados utilizada contém notícias de três cadernos - Dinheiro, Emprego, Esportes - de edições do jornal Folha de São Paulo publicadas entre 1993 e 1994. A ferramenta PreText 2 [16] foi utilizada para pré-processamento dos textos. Símbolos, números e pontuação foram removidos, bem como *stopwords*[4]. Por fim, foi obtida uma matriz binária contendo 4828 termos e 300 notícias (100 de cada caderno).

Os algoritmos NMF e NBVD, descritos na seção 3, foram aplicados à tarefa de agrupamento das notícias. O objetivo é identificar quais notícias pertencem a quais cadernos com base somente nos seus conteúdos textuais. Para comparação, o algoritmo *k-means* foi aplicado à mesma tarefa. Os resultados produzidos pelos algoritmos foram medidos de acordo com índice de avaliação externo MAP (*micro-averaged precision*) [7], que fornece a acurácia em relação à correta atribuição de notícias aos respectivos cadernos. Vale ressaltar que no contexto deste trabalho, os rótulos associados a cada documento são utilizados apenas como conhecimento externo para avaliação. Os algoritmos utilizados não tem conhecimento *a priori*, tampouco o processo de recomendação proposto beneficiar-se-á desta informação.

A acurácia média obtida pelo *k-means* foi de $0,39$ ao passo que NMF e BVD obtiveram, respectivamente, $0,67$ e $0,63$. As acurácias reportadas correspondem à média de 30 execuções. Todos os resultados apresentaram desvio padrão abaixo de $0,06$. Os resultados indicam que os algoritmos de coagrupamento mantém razoável acurácia nas predições,

[4]Palavras com baixo poder de discriminação entre textos. Geralmente, pronomes, artigos e conjunções.

o que os credenciam como possíveis algoritmos para oferecer recomendações com mínimo grau de relevância. Além disso, o coagrupamento de dados apresenta como diferencial a capacidade de agrupar dados em ambas as dimensões simultaneamente, e.g. termos e notícias, o que possibilita a identificação de grupos de termos que contenham informações implícitas das relações entre os itens. Uma análise exploratória desses grupos, incluindo a parametrização do algoritmo de coagrupamento para criação de diferentes configurações (número) de grupos de termos, pode ser o caminho para a construção de recomendações serendipitosas.

6. CONCLUSÃO

Neste artigo o problema da serendipidade em sistemas de recomendação baseados em conteúdo foi formalmente definido e uma proposta de solução apresentada. O coagrupamento de dados é peça fundamental na solução proposta, pois permite a identificação de grupos de itens parcialmente similares. Experimentos preliminares sugerem que os algoritmos de coagrupamento são capazes de manter níveis razoáveis de acurácia no agrupamento de documentos e, ao mesmo tempo, fornecem informações adicionais a respeito do agrupamento de termos que podem ser exploradas em busca de recomendações serendipitosas. Assim, os resultados apresentados indicam o potencial do coagrupamento de dados para solução do problema da serendipidade. Próximos passos incluem a aplicação de outras operações de pré-processamento e representação dos dados como *stemming*, *tf-idf*, normalizações e suavizações, a incorporação do resultado do coagrupamento em uma estratégia de recomendação e avaliação do sistema de recomendação.

7. REFERÊNCIAS

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

[2] M. Balabanovic and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, Mar. 1997.

[3] J. Bennett, S. Lanning, et al. The netflix prize. In *Proceedings of KDD Cup and workshop*, volume 2007, page 35. New York, NY, USA, 2007.

[4] P. A. Castro, F. O. França, H. M. Ferreira, and F. J. Von Zuben. Applying biclustering to perform collaborative filtering. In *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on*, pages 421–426. IEEE, 2007.

[5] Y. S. Chiu, K. H. Lin, and J. S. Chen. A social network-based serendipity recommender system. In *Intelligent Signal Processing and Communications Systems (ISPACS), 2011 International Symposium on*, pages 1–5. IEEE, 2011.

[6] M. de Gemmis, P. Lops, G. Semeraro, and C. Musto. An investigation on the serendipity problem in recommender systems. *Information Processing & Management*, 51(5):695–717, 2015.

[7] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 89–98. ACM, 2003.

[8] M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the 4th ACM Conference on Recommender Systems*, pages 257–260. ACM, 2010.

[9] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 17(2):107–145, 2001.

[10] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[11] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering. In *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*. 2007.

[12] G. Linden, B. Smith, and J. York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[13] B. Long, Z. M. Zhang, and P. S. Yu. Co-clustering by block value decomposition. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 635–640. ACM, 2005.

[14] M. G. Manzato and R. Goularte. A multimedia recommender system based on enriched user profiles. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 975–980. ACM, 2012.

[15] A. L. V. Pereira and E. R. Hruschka. Simultaneous co-clustering and learning to address the cold start problem in recommender systems. *Knowledge-Based Systems*, 82:11–19, 2015.

[16] M. V. B. Soares, R. C. Prati, and M. C. Monard. *Pretext II: Descriçao da reestruturacao da ferramenta de pré-processamento de textos*. ICMC-USP, 2008.

[17] D. D. Woods, E. S. Patterson, and E. M. Roth. Can we ever escape from data overload? a cognitive systems diagnosis. *Cognition, Technology & Work*, 4(1):22–36, 2002.

[18] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–273. ACM, 2003.

[19] H. Yamaba, M. Tanoue, K. Takatsuka, N. Okazaki, and S. Tomita. On a serendipity-oriented recommender system based on folksonomy. *Artificial Life and Robotics*, 18(1-2):89–94, 2013.

[20] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32. ACM, 2005.

Ensemble de agrupamentos para recomendações serendipitosas baseada em conteúdo

Cluster ensemble to content-based serendipitous recommendations

Fernando Henrique da Silva Costa
Universidade de São Paulo
Av. Arlindo Bettio, 1000
São Paulo, SP, Brasil
fhscosta0993@usp.br

Sarajane Marques Peres
Universidade de São Paulo
Av. Arlindo Bettio, 1000
São Paulo, SP, Brasil
sarajane@usp.br

RESUMO

Sistemas de recomendação são ferramentas capazes de recomendar aos usuários itens relevantes e personalizados. Esses sistemas são divididos em arquiteturas, sendo uma delas a arquitetura de recomendação baseada em conteúdo. Tal arquitetura recomenda itens ao usuário com base na similaridade entre os itens e nas avaliações que usuários fizeram sobre itens no passado. Uma limitação é que ela, geralmente, faz recomendações de baixa serendipidade, visto que os itens recomendados são parecidos com os itens já observados pelo usuário, e não representam novidade, diversidade ou surpresa. Esta pesquisa objetiva minimizar o problema da falta de serendipidade fazendo uso de análise de similaridades parciais. A fim de apresentar uma solução eficiente em termos de custo, essa análise será implementada a partir de *ensembles* de agrupamento. O resultado do *ensemble* será usado como base para construção das recomendações. Experimentos no contexto de notícias serão realizados como prova de conceito da abordagem proposta.

Palavras-Chave

Ensemble de agrupamento, sistemas de recomendação baseados em conteúdo, serendipidade, similaridade parcial.

ABSTRACT

Recommender systems are tools able of recommending relevant and personalized items to users. These systems are divided into architectures, one of which is the content-based recommendation architecture. Such architecture recommends items based on the similarity between itens and the evaluations that users have made about items in the past. A limitation is that such arquitecture generally makes recommendations of low serendipity, since the recommended items are similar to the items already observed by the user, and do not represent novelty, diversity or surprise. This research aims to minimize the problem of lack of serendipity by using partial similarities analysis. Also, in order to provide an efficient solution in terms of computational cost, this analysis will be implemented through cluster ensembles. The result of the ensemble will be used as a basis for building recommendations. Experiments with news recommendations will be held as proof of concept of the proposed approach.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•Information systems → Clustering; Content analysis and feature selection; Personalization; •Computing methodologies → Cluster analysis;

Keywords

Cluster ensemble, content-based recommender systems, serendipity, partial similarity

1. INTRODUÇÃO

O crescimento de produções de texto na internet gerou uma quantidade considerável de informações, além de uma alta disponibilidade para elas. Entretanto, tal volume de informação originou algumas dificuldades. Por exemplo, um usuário que possui pouca ou nenhuma experiência para escolha de uma alternativa dentre as várias apresentadas terá dificuldade em encontrar itens úteis e que atendam às suas necessidades [14]. Uma solução para resolver tais dificuldades é a proposição de um mecanismo que seja capaz de recomendar itens relevantes e úteis ao usuário. Nesse contexto surgem os sistemas de recomendação, os quais são ferramentas de *software* capazes de sugerir itens relevantes, por exemplo, sugerir quais músicas ouvir ou quais notícias ler [16]. Hoje, tais sistemas são capazes de produzir recomendações individualizadas ou têm o propósito de orientar o usuário de forma personalizada em um espaço grande de opções [4]. As recomendações individuais são fornecidas na forma de listas de itens ordenados, que são construídas como uma tentativa de predizer os itens mais adequados ao usuário tendo em vista suas preferências e restrições [16].

No contexto de sistemas de recomendação há várias arquiteturas, sendo as principais baseadas em: em conteúdo, em filtros colaborativos e em conhecimento. Este trabalho estuda a primeira arquitetura, a qual recomenda itens similares aos itens que o usuário já tenha gostado no passado [1]. Os sistemas de recomendação baseado em conteúdo analisam um conjunto de itens, ou de descrições de itens, avaliados previamente pelo usuário e que, assim, constituem o seu

perfil de interesse [14]. Há vários cenários de aplicação para um sistema de recomendação e, dentre eles, existem aqueles que recomendam músicas, filmes e notícias. Neste trabalho, o cenário de interesse é o de notícias publicadas em portais, portanto, trata-se de uma recomendação baseada em conteúdo textual. Os sistemas de recomendação que atuam nesse cenário estão inseridos em um ambiente crítico de serviços *online*. Neste cenário há um volume de informações com o qual o usuário tem dificuldade de lidar, sentindo-se desconfortável [13]. Portanto, um sistema de recomendação pode ajudar a encontrar notícias pertinentes.

2. APRESENTAÇÃO DO PROBLEMA

Inicialmente, os sistemas de recomendação buscavam alcançar a relevância da recomendação [15]. Conforme esse princípio, a recomendação de itens adequados à preferência do usuário contribuiria para a satisfação do mesmo. Todavia, as recomendações que visam a alta acurácia não atendem a outras qualidades pertinentes à uma recomendação e, assim, não atendem as expectativas do usuário. O problema de manter o foco em recomendações com alta acurácia está presente na arquitetura baseada em conteúdo. Contudo, ela possui algumas vantagens, como a independência de outros usuários para a geração de recomendação (que é um problema na arquitetura de filtro colaborativo), a transparência de funcionamento do sistema e a independência de avaliações sobre novos itens inseridos no sistema, sendo possível realizar as recomendações apenas com base nos atributos do item (solução para o problema de *cold start* de item) [14].

Buscando melhorar a qualidade da recomendação, novos aspectos de avaliação em sistema de recomendação foram elaborados, e exemplos são os aspectos de novidade, diversidade e serendipidade. O primeiro aspecto diz respeito a itens que o usuário não tenha conhecimento, ainda que tais itens sejam totalmente compatíveis com o seu perfil de interesse. O segundo aspecto é o oposto de similaridade, ou seja, são os itens que estão fora do escopo de interesse do usuário. O último aspecto determina o quão surpreendente são as recomendações bem sucedidas [17]. Uma recomendação serendipitosa ajuda um usuário a encontrar itens surpreendentes e interessantes, que não seriam descobertos por outro meio [8]. Para ilustrar a diferença entre serendipidade e novidade, considere uma recomendação que diz respeito à filmes do diretor favorito do usuário, porém que não foram assistidos por ele. Esta recomendação será nova, mas sem surpresa ou de baixa serendipidade. Todavia, a recomendação de um filme de um diretor desconhecido pelo usuário, porém que versa sobre um tema de seu interesse, tem maior probabilidade de ser serendipitosa.

O problema estudado nesta pesquisa, que está associado a uma limitação da arquitetura baseada em conteúdo, é o problema de recomendações de baixa serendipidade. Este problema está relacionado a um sistema que recomenda apenas itens similares à outros que o usuário já teve interesse [3], sendo limitado em recomendar itens novos e surpreendentes.

3. PROPOSTA DA SOLUÇÃO

A solução proposta neste trabalho consiste no uso de um *ensemble* de agrupamento como base para as recomendações. A ideia é usar como base o que é discutido nos trabalhos apresentados em [5, 18]. Tais trabalhos utilizam, respectivamente, técnicas de agrupamento como um método baseado em modelo, e *ensemble* de agrupamento em sistemas de recomendação de filtragem colaborativa.

Um *ensemble* de agrupamentos visa melhorar a qualidade do agrupamento gerando várias partições de um conjunto de dados e combinando-as para formar uma solução final [7]. A motivação para a escolha em usar *ensemble* de agrupamentos está pautada em duas questões: a primeira delas é a possibilidade de implementar uma solução de agrupamento baseada na análise de similaridades parciais entre itens, o que possibilitaria a construção de recomendações com alguma surpresa para o usuário; a segunda é necessidade de obter uma base de recomendação que tenha uma boa acurácia [10], de forma a propiciar que as recomendações tenham relevância. Enfim, espera-se atender aos dois aspectos que compõem a serendipidade: a surpresa e a relevância.

O escopo de trabalho da recomendação esperada, a partir da estratégia proposta, é mostrado na figura 1. Essa figura apresenta na primeira parte o espectro de recomendação divida nas características de surpresa e alta acurácia. A surpresa é dividida em duas partes: negativa e positiva. É esperado que a estratégia de recomendação proposta neste trabalho seja capaz de recomendar itens referentes a segunda parte da surpresa, ou seja, itens que são surpreendentes e que possuem características que estão de acordo com o interesse do usuário e são, portanto, uteis a ele. O asterisco na figura mostra a área em que é esperado a atuação do sistema de recomendação.

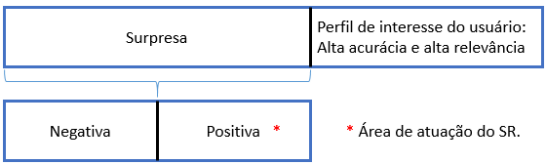


Figura 1: Atuação do sistema de recomendação

A fim de esclarecer como a similaridade parcial será implementada no *ensemble*, considere que o conjunto de dados a ser agrupado no *ensemble*, ou seja o conteúdo textual das notícias, será representado por um vetor de atributos [6]. Em uma estratégia de agrupamento clássica, todo esse vetor de atributos seria apresentado ao algoritmo para que esse executasse o processo de agrupamento considerando a similaridade total entre as notícias. Usando um *ensemble* de agrupamento é possível estruturar cada um de seus componentes a partir de um subconjunto de atributos da notícia, dessa forma, cada componente estaria analisando a similaridade entre as notícias sob um aspecto diferenciado e cada um deles estaria analisando similaridades parciais.

A construção de grupos usando a ideia de similaridades parciais é bastante difundida nos algoritmos de coagrupamento, os quais descobrem "blocos" de linhas e colunas que estão correlacionados [9]. No entanto, tais algoritmos são de alto custo computacional e, considerando que textos são tipos de dados representados em alta dimensionalidade, o uso deles podem representar uma dificuldade à solução. A fim de superar essa dificuldade, a presente pesquisa propõe o uso dos *ensembles* de agrupamento como uma forma de obter um resultado equivalente ao de coagrupamento.

O objetivo em utilizar *ensemble* de agrupamento sob a perspectiva da análise de similaridades parciais está relacionado a intenção de implementar uma estratégia de extração

de informação das notícias. Uma vez que é esperado que os blocos de linhas e colunas correlacionados sejam encontrados, é possível extrair uma informação mais específica desses blocos em relação aos atributos que fazem determinadas notícias semelhantes, e prover uma justificativa para as recomendações. A hipótese defendida neste trabalho é que a base de recomendação pretendida propiciará resultados mais relevantes em termos de serendipidade quando comparados aos resultados que estão apresentados na literatura da área.

4. AVALIAÇÃO DA SOLUÇÃO

Recomendações com alta relevância podem não ser serendipitosas e as muito diversificadas podem ser irrelevantes. É esperado, em vista da solução proposta nesta pesquisa, que o *ensemble* de agrupamento seja capaz de criar uma lista de recomendações relevantes e surpreendentes. Normalmente, para a avaliação da relevância dos resultados produzidos em agrupamento são usados índices internos ou externos. Para esta pesquisa, os índices internos escolhidos para a aplicação são Silhouette e Dunn; os índices externos são Rand e Informação Mútua Normalizada. Enquanto os índices externos avaliam os resultados apoiados por uma estrutura especificada *a priori*, os índices internos avaliam a densidade de agrupamento e separabilidade de grupos. A avaliação da acurácia do resultado do *ensemble* vai propiciar a avaliação da relevância de uma recomendação baseada nesse *ensemble*. Uma vez que possua um *ensemble* que constrói grupos de notícias baseada na similaridade parcial e possui uma acurácia aceitável, ele será utilizado para realizar recomendações, que serão avaliadas em termos de medidas de surpresa [12]. Duas medidas de surpresa são apresentadas em [11] seguindo a ideia de medir a serendipidade de um item a partir da sua distância até aos itens que o usuário já observou.

Para que a solução proposta seja avaliada em sua completude será necessário a realização de experimentos. As formas de realizar experimentos no contexto de sistemas de recomendação são *online*, *offline* e *user study*. A fim de medir de forma qualitativa os resultados produzidos nesse trabalho, o experimento *user study* será usado. No experimento será solicitado a um conjunto de usuários executar tarefas usando um sistema de recomendação. Tipicamente, as tarefas são a navegação no sistema e a resposta à questionários da experiência [17] de navegação analisando a serendipidade.

5. ATIVIDADES JÁ REALIZADAS

Até o momento foram realizados levantamentos teóricos sobre comitê de máquinas, *ensemble* de agrupamento e sistemas de recomendação baseado em conteúdo. Atualmente, está em fase de finalização uma revisão sistemática sobre *ensembles* de agrupamento aplicados à conteúdos textuais. Essa revisão visa levantar quais são as principais técnicas usada em *ensembles* de agrupamento, como é realizada a geração de componentes, quais são as funções de consenso disponíveis, quais métricas de similaridade são usadas e como os resultados são avaliados. Outra revisão sistemática está sendo realizada no âmbito do grupo de pesquisa, a qual versa sobre a característica de serendipidade. O objetivo dessa revisão é levantar quais são as técnicas comumente aplicadas em sistemas de recomendação que recomendam com alta serendipidade, e como eles são avaliados. Por fim, está sendo realizado um estudo exploratório em *ensemble* para sistemas de recomendação. Como resultado desses estudos, foi observado que para a geração dos componentes de um *ensemble* costuma-se aplicar algoritmos que são de execução simples, como o *k-means*. Quanto a métrica de similaridade, desde que a pesquisa está sendo desenvolvida sobre conteúdo textual, a métrica mais difundida é a distância cosseno. E a abordagem para consenso mais citada é a matriz de coassociação. Por fim, as avaliações são geralmente feitas por medidas de *F-measure* e Informação Mútua Normalizada.

A partir de constatações obtidas no estudo literatura, foi decidido iniciar os testes a partir da implementação de um *ensemble* de agrupamento usando *k-means++* [2]. As distâncias euclidiana e cosseno foram implementadas a fim de realizar uma comparação entre o efeito das mesmas sobre os dados textuais. Para a função de consenso foi usada a matriz de coassociação e um algoritmo de agrupamento hierárquico aglomerativo com método de *linkage*. O *ensemble* foi aplicado na base de notícias Folha 93-94, pré-processada com a ferramenta PreText e representada via *tf-idf* não normalizado. Futuramente será usada a base de notícias oriundas do portal EBC (Empresa Brasil de Comunicação).

6. CONCLUSÃO

A proposta elaborada nesta pesquisa tem como objetivo construir uma base de recomendação por meio da utilização de um algoritmo de *ensemble* de agrupamento projetado de maneira a implementar a similaridade parcial entre itens. Esta similaridade parcial será obtida com a aplicação dos componentes do *ensemble* em conjuntos de dados formados por subconjuntos de atributos do conjunto de dados original, como uma estratégia de extração de informações de notícias.

Os próximos passos para completar a execução desta pesquisa são: finalizar as revisões sistemáticas e exploratória, preparar a experimentação com os usuários, estudar estratégias de função de consenso que não priorizem somente a acurácia, mas que permita a partição final conter as características esperadas de surpresa oriundas das similaridades parciais, avaliar os resultados para construir o mecanismo de recomendação que será utilizado no sistema e, por fim, a análise dos dados coletados da experiência do usuário.

7. REFERÊNCIAS

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.

[2] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[3] M. Balabanović and Y. Shoham. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[4] R. Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[5] F. Darvishi-Mirshekarlou, S. Akbarpour, M. Feizi-Derakhshi, et al. Reviewing cluster based collaborative filtering approaches. *International*

Journal of Computer Applications Technology and Research, 2(6):650–meta, 2013.

[6] F. O. de França. *Biclusterização na análise de dados incertos*. PhD thesis, Universidade Estadual de Campinas, 2010.

[7] X. Z. Fern and C. E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the twenty-first international conference on Machine learning*, page 36. ACM, 2004.

[8] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[9] S. Huang, H. Wang, D. Li, Y. Yang, and T. Li. Spectral co-clustering ensemble. *Knowledge-Based Systems*, 84:46–55, 2015.

[10] N. Iam-On, T. Boongoen, S. Garrett, and C. Price. A link-based approach to the cluster ensemble problem. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2396–2409, 2011.

[11] M. Kaminskas and D. Bridge. Measuring surprise in recommender systems. In *Proceedings of the Workshop on Recommender Systems Evaluation: Dimensions and Design (Workshop Programme of the 8th ACM Conference on Recommender Systems)*. Citeseer, 2014.

[12] D. Kotkov, S. Wang, and J. Veijalainen. A survey of serendipity in recommender systems. *Knowledge-Based Systems*, 111:180–192, 2016.

[13] J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 31–40. ACM, 2010.

[14] P. Lops, M. De Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender Sytems Handbook*, pages 73–105. Springer, 2011.

[15] K. Oku and F. Hattori. Fusion-based recommender system for improving serendipity. In *DiveRS@ RecSys*, pages 19–26, 2011.

[16] F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[17] G. Shani and A. Gunawardana. Evaluating recommendation systems. In *Recommender Systems Handbook*, pages 257–297. Springer, 2011.

[18] C.-F. Tsai and C. Hung. Cluster ensembles in collaborative filtering recommendation. *Applied Soft Computing*, 12(4):1417–1425, 2012.

Sistema de Alerta Antecipado Sensível ao Contexto: Uma Abordagem Baseada em Textos para Cegos e Surdos

Alternative Title: Context-Aware Early Warning System: A text-based approach for blind and deaf

Jaziel Souza Lobo
Instituto Federal de Sergipe /
Universidade Federal da Bahia
Salvador, Bahia
jaziel.lobo@ifs.edu.br

Vaninha Vieira
Universidade Federal da Bahia
Departamento de Ciência da Computação
Salvador, Bahia
vaninha@ufba.br

RESUMO

Desastres naturais ou provocados pelo homem podem acontecer de forma inesperada e levar à perda de vidas e danos à sociedade. A fim de alertar a população, os sistemas de alerta precoce enviam mensagens sobre situações que ainda vão acontecer. De acordo com um estudo de mapeamento sistemático conduzido pelos autores, há pouca pesquisa no campo dos sistemas de alerta precoce focados em populações vulneráveis como cegos e surdos. As instituições de previsão, como os institutos meteorológicos, fornecem alertas sobre situações futuras no formato CAP. O padrão CAP é projetado para trocar mensagens de alerta entre sistemas. As informações CAP são estáticas e não permitem a adaptação para as populações vulneráveis. Este trabalho propõe o desenvolvimento de um sistema de alerta precoce com foco na adaptação de textos de mensagens de alerta no padrão CAP para cegos e surdos.

Palavras-Chave

Sistema de alerta antecipado; Sistema sensível ao contexto; Comunicação de crise; População vulnerável; Cego; Surdo.

ABSTRACT

Natural or man-made disasters can happen in an unexpected way and lead to loss of life and damage to the society. In order to alert the population, early warning systems send messages about situations that are still going to happen. According to a systematic mapping study conducted by the authors, there is little research in the field of early warning systems focused on vulnerable populations as the blind and deaf. Prediction institutions such as meteorological institutes provide alerts on future situations in the CAP format. The CAP standard is designed to exchange alert messages between systems. CAP information are static and does not allow the adaptation to vulnerable populations. This work proposes the development of an early warning system focusing on the adaptation of texts of alert messages in the CAP standard for blind and deaf people.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•Human-centered computing → Ubiquitous and mobile computing systems and tools;

Keywords

Early warning system; Context-aware system; Crisis Communication; Vulnerable population; Blind; Deaf

1. INTRODUÇÃO

Ao longo dos anos desastres, sejam eles naturais ou provocados pelo homem, têm afetado a vida das pessoas ao redor do mundo. O relatório anual de 2009 da Cruz Vermelha Americana mostra que muitos dos feridos fazem parte de populações vulneráveis como idosos e pessoas com algum tipo de deficiência [9]. Um desastre é definido como um evento que causa uma interrupção séria no funcionamento da sociedade que leva a grandes perdas humanas, materiais e ambientais, e a sociedade afetada é incapaz de lidar com a situação utilizando seus próprios recursos [4].

Para Wurster e Meissen [17], uma forma de tentar minimizar o número de vítimas e danos é alertar antecipadamente a população. Antigamente os alertas eram transmitidos através de rádio e TV e mais recentemente os alertas estão sendo transmitidos por "Sistema de Alerta Antecipado"através de SMS, e-mail ou das redes sociais [2]. Os Sistema de Alerta Antecipado, ou EWS (*Early Warning Systems*), são sistemas com o conjunto de capacidades necessárias para gerar e disseminar avisos em tempo suficiente para que indivíduos, comunidades e organizações ameaçadas por um perigo se preparem e ajam adequadamente para reduzir a possibilidade de danos ou perdas [14].

EWS podem ser classificados em dois tipos: Sistemas Baseados em Radiodifusão e Sistemas Baseados em Assinatura. Sistemas baseados em radiodifusão são caracterizados por enviar alertas para todos os destinatários potenciais sem qualquer registro prévio, bastando que eles estejam em área de risco. Sistemas baseados em assinatura exigem que os destinatários realizem um cadastro prévio para receber aler-

tas de desastre[6]. A maioria dos sistemas de alerta antecipado utilizam métodos de disseminação massiva de mensagens, como SMS ou Cell-Broadcasting [8]. Soluções de divulgação em massa enviam a mesma mensagem para diferentes tipos de pessoas ou público alvo, como possíveis vítimas, parentes das vítimas, pessoas com deficiência, idosos, imprensa, etc [8] [15].

De acordo com Sullivan e Häkkinen [13], as populações vulneráveis como os deficientes ou idosos estão sujeitas a um risco especial em um desastre e é fundamental considerar as suas necessidades especiais na concepção de sistemas de preparação e aviso de catástrofes. Os sistemas baseados em radiodifusão disseminam a mesma mensagem para todos, sem considerar qualquer necessidade especial, entretanto, é preciso que as mensagens de alerta considerem as necessidades especiais de cada indivíduo. O primeiro passo para enviar mensagens individualizadas é entender o contexto da situação e da pessoa. Para Dhokane et al. [5], contexto é qualquer informação que pode ser usada para caracterizar a situação de entidades (uma pessoa, um lugar ou um objeto) e que são consideradas relevantes para a interação entre um usuário e uma aplicação, incluindo o próprio usuário e aplicação em si. De acordo com Vieira et al. [16], "em um sentido amplo, o contexto é tudo o que envolve uma situação, em um dado momento, e que permite identificar o que é ou não relevante para interpretar e entender essa situação". Dessa forma, para que um sistema de alerta envie mensagens individualizadas, é preciso que ele seja sensível ao contexto. Sistema sensível ao contexto é a classificação dada a todos os sistemas que utilizam o contexto para fornecer serviços ou informações mais relevantes para apoiar os utilizadores que executam as suas tarefas [16].

Diante disto, este trabalho apresenta a proposta de um Sistema de Alerta Antecipado Sensível ao Contexto para o domínio de crise e emergência com foco na população vulnerável de deficientes visuais e auditivos. Para discorrer sobre o assunto, este artigo foi organizado da seguinte forma: Na Seção 2 apresentamos o problema e na Seção 3 a solução proposta para resolver o problema em questão. A Seção 4 descreve o projeto para a avaliação da solução. Na Seção 5 são relatadas as atividades já realizadas e, para finalizar, as conclusões são apresentadas na Seção 6.

2. APRESENTAÇÃO DO PROBLEMA

Segundo Phillips e Morrow [12], apenas transmitir aviso para alertar a população é insuficiente, pois o conteúdo da mensagem vai ser interpretado dentro do contexto sociocultural do receptor. Esta constatação pode ser reforçada através dos relatos encontrados na literatura cujo conteúdo da mensagem de alerta influenciou negativamente para o receptor. Como exemplo pode ser citado o tornado em Saragosa no Texas em 1987. Naquela ocasião os meios de comunicação locais distorceram a tradução da palavra "warning"do inglês para o espanhol e vinte e nove vidas foram perdidas porque as estações de televisão em língua espanhola que estavam fora da área local eram proibidas de transmitir avisos locais [12]. Também existem relatos sobre a dificuldade de compreensão de textos por pessoas surdas: Entrevistado 1: "Às vezes é bem complicado para os surdos compreenderem as leituras, os textos, acabo passando por certa dificuldade, em diferentes contextos sofro, mas luto para sobreviver, tenho que estudar, bem faço com esforço.". Entrevistado 3: "Se é uma leitura fácil, então é fácil de compreender, mas se é uma leitura difícil, não tenho conhecimento do vocabulário, eu vou ter que pesquisar no dicionário para conseguir entender o significado destas palavras."[3].

Estudos sobre o desenvolvimento de jovens surdos identificaram que existe um baixo nível de letramento e dificuldades na aprendizagem da leitura e escrita de textos em português [3]. Os deficientes visuais compõem outro grupo que faz parte da população vulnerável. Para este grupo, existem aplicativos e sistemas operacionais de smartphones que convertem texto para voz através de uma tecnologia conhecida como TTS (*Text-To-Speech*). Apesar da tecnologia ajudar neste aspecto, pouco se sabe a respeito de como os deficientes visuais preferem receber previsões e avisos de desastres ou como eles interpretam a informação recebida.

Na tentativa de criar um padrão para transmissão de mensagens de alerta, o Conselho Nacional de Ciência e Tecnologia dos Estados Unidos (*National Science and Technology Council*), em novembro de 2000, fez uma recomendação sobre a necessidade de desenvolver um método padrão para coletar e retransmitir avisos de perigo para sistemas de disseminação de alertas [10]. Em resposta a essa recomendação, um grupo de trabalho internacional começou a desenvolver o Protocolo de Alerta Comum ou CAP como é amplamente conhecido [11]. Este grupo era formado por mais de 130 gestores de emergência e especialistas em tecnologia da informação e telecomunicações. A estrutura de uma mensagem CAP descreve um evento real ou a sua previsão, com detalhes sobre sua urgência, gravidade e área geográfica atingida. Além dessas informações, mensagens CAP podem fornecer aos destinatários instruções de como agir e outras informações como: duração do perigo, parâmetros técnicos, informações de contato, links para fontes de informação adicionais, etc.

Apesar de contemplar diversas informações, os dados trafegados em mensagens CAP são todos estáticas e não fazem referência à população vulnerável. Para preencher essa lacuna, alguns trabalhos utilizaram a sensibilidade ao contexto com diferentes finalidades para adaptação. Com o objetivo de identificar trabalhos que utilizaram sensibilidade ao contexto em sistemas de alerta antecipado, realizamos em 2015 um mapeamento sistemático da literatura. A seguir destacamos alguns destes trabalhos.

Malizia et al. [7] apresenta um protótipo que envia notificações de emergência personalizadas de forma automática. O protótipo usa uma ontologia chamada SEMA4A. A ontologia fornece uma representação e a relação em domínios de acessibilidade, mídia e dispositivo. Através de consultas realizadas na ontologia SEMA4A, o protótipo extrai as possíveis formas de enviar o alerta, por exemplo: dispositivo (TV, rádio, telefone celular, telefone) e mídia (vídeo, som, figura, texto, vibração).

Mitchell et al. [9] apresenta um protótipo de alertas de emergência acessível. O estudo avaliou o envio de alertas para pessoas surdas e com deficiência auditiva. As mensagens eram enviadas em texto e em Linguagem de Sinais Americana (ASL). Os resultados mostraram que os participantes consideram que os alertas de vídeo ASL representavam uma ferramenta útil para as pessoas surdas. Por outro lado, criticaram a qualidade da interpretação ASL e classificaram como inadequadas algumas escolhas de vocabulário e algumas frases traduzidas para ASL. Esta abordagem torna evidente que não é suficiente enviar alertas por diferentes formatos. O conteúdo da mensagem precisa ser trabalhado

para que o público-alvo possa compreendê-la.

Aedo et al. [1] propõe um conjunto de critérios para adaptar alertas de emergência e vias de evacuação para diferentes situações e pessoas em ambientes fechados. A adaptação ocorre de acordo com as preferências dos usuários, dispositivos e tipo de deficiência. A proposta também utiliza a ontologia SEMA4A. Através da ontologia, o sistema de alerta escolhe o formato e o meio para enviar a mensagem, como SMS, MMS, e-mail e notificações por aplicativos no smartphone. Nesta abordagem, o texto é enviado da mesma forma através de diferentes meios.

Klafft and Ziegler [6] propõem uma arquitetura para a integração de sistemas de alerta de desastre. A arquitetura oferece a possibilidade de personalização de mensagens de alerta para diferentes grupos de destinatários, como turistas, pessoas com deficiência, famílias com crianças. Em um exemplo, as famílias com crianças foram recomendadas a buscar seus filhos na escola. Em outro, famílias com crianças foram recomendados a não buscar os filhos na escola e abrigarem-se em casa. Na arquitetura proposta as mensagens são enviadas para todos da mesma forma, apenas incluindo ou excluindo textos de acordo com a situação de cada grupo. Mas não existe qualquer adaptação do conteúdo da mensagem voltado para pessoas com deficiência.

Estes estudos sugerem adaptação da mensagem para indivíduos com deficiência, no entanto, a personalização é restrita à definição dos meios de comunicação e adaptação ao dispositivo receptor. Diante da importância na personalização das mensagens para o usuário, identificou-se como problema a deficiência dos sistemas de alerta antecipado do domínio de crise e emergência para o envio de avisos de texto com foco para a população vulnerável de deficientes visuais e auditivos.

3. PROPOSTA DE SOLUÇÃO

Este trabalho propõe desenvolver um sistema de alerta antecipado com foco na adaptação dos textos das mensagens CAP para pessoas cegas e surdas que estão em área de risco. Através desta adaptação, pretende-se aumentar a compreensão das mensagens enviadas para os usuários. A arquitetura conceitual proposta para o trabalho é apresentada na Figura 1. Como pode ser observado, a arquitetura possui três componentes: Gerenciador de Contexto (GC), Construtor de Mensagens (CM) e Disseminador.

O Gerenciador de Contexto será responsável pela aquisição, interpretação e raciocínio das informações contextuais. Nesse primeiro momento propõe-se um modelo de contexto contemplando três entidades contextuais: (i) emergência, (ii) usuário e (iii) dispositivo. Os elementos contextuais a serem adquiridas para o item (i) serão dados sobre a situações de emergência que estão especificados no padrão CAP. Os elementos contextuais para o item (ii) serão coletados através de um perfil e deverá contemplar o tipo de deficiência, a idade e grau de escolaridade. Para o item (iii) serão coletados dados do dispositivo receptor como a localização através de coordenadas GPS.

O "Construtor de Mensagens", terá a responsabilidade de adaptar o conteúdo das mensagens de acordo com as informações recebidas do GC. Dessa forma, é preciso que esse componente identifique os prováveis usuários afetados pela situação e o tipo de adaptação que deverá ser realizada. A identificação dos usuários afetados deverá ser realizada com base na sua localização e na área de risco do desastre. Com

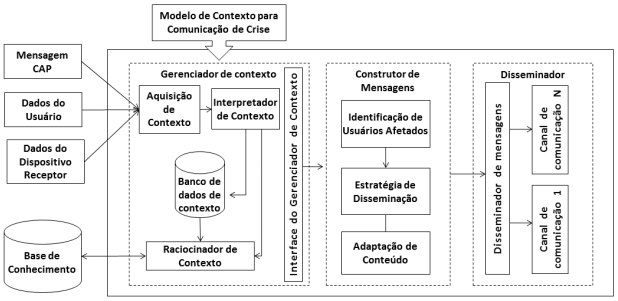


Figura 1: Arquitetura conceitual preliminar.

relação a adaptação de conteúdo da mensagem, serão utilizadas as informações do perfil do usuário como o tipo de deficiência, sua faixa etária e o nível de escolaridade. Com base ainda nas informações contextuais do GC, o CM definirá qual a melhor estratégia para disseminação dos alertas. Finalmente a transmissão das mensagens de alerta será realizada pelo componente Disseminador.

Para aperfeiçoar a elaboração do modelo de contexto, serão realizadas entrevistas com deficientes auditivos e deficientes cegos para identificar as dificuldades que este público possui no entendimento de mensagens do domínio de crise e emergência. Nas análises dos dados serão procuradas correlações das informações dadas com variáveis como faixa etária e grau de escolaridade para definição dos elementos contextuais relevantes para o modelo.

4. PROPOSTA DE AVALIAÇÃO

Pretende-se avaliar a solução proposta através de duas perspectivas: (i) avaliar se as mensagens enviadas foram entregues aos destinatários corretos e com a personalização exata; e (ii) avaliar se a adaptação utilizada ajudou os diferentes tipos de usuários. Para a avaliação (i), será utilizada uma abordagem empírica através de técnicas quantitativas. As técnicas quantitativas serão aplicadas para a verificação da precisão na entrega das mensagens, ou seja, se elas foram entregues aos destinatários corretos e com a personalização exata. Para a avaliação (ii), será planejado e executado um quasi-experimento onde participantes, dos dois públicos, formarão dois agrupamentos. O primeiro grupo receberá mensagens disseminadas em massa com o texto recebido diretamente na mensagem CAP. No segundo grupo as mensagens serão enviadas de forma adaptada, de acordo com personalização da proposta. Através de entrevistas poderá ser constatado se as mensagens entregues foram relevantes ou não para o usuário.

5. ATIVIDADES JÁ REALIZADAS

Com o objetivo de identificar *gaps* na área de comunicação de crise no domínio de crise e emergência, um mapeamento sistemático da literatura foi realizado. No mapeamento as questões de pesquisa foram focadas para: (i) analisar o uso da consciência de contexto; (ii) Analisar as situações de desastre cobertas; (iii) caracterizar mensagens enviadas; (iv) Analisar o apoio às pessoas com deficiência. Os resultados apontaram que existe um número reduzido de trabalhos para pessoas com deficiência no domínio de crise e emergência. Outra atividade realizada foi a condução uma entrevista

preliminar com algumas pessoas que fazem parte da Associação Baiana de Cegos. O objetivo foi descobrir melhores formas de conduzir entrevistas para este público e conhecer se eles já tinham se envolvido em situações de desastres e como eles tinham sido alertados.

6. CONCLUSÃO

A proposta deste trabalho é desenvolver um sistema de alerta antecipado com foco na adaptação dos textos das mensagens CAP para pessoas cegas e surdas que estão em área de risco. Através desta adaptação, pretende-se melhorar o entendimento das mensagens de alerta pelos usuários. O resultado desse trabalho trará contribuições científicas e tecnológicas. Nas contribuições científicas, pretende-se somar ao corpo de conhecimento (i) um modelo de contexto do domínio de crise e emergência para ser utilizado por sistemas de alerta antecipado na comunicação de crises; (ii) estratégias de adaptação de textos de mensagens sobre desastres para alertar cegos e surdos. Como contribuição tecnológica, teremos o desenvolvimento de um sistema de alerta antecipado sensível ao contexto para comunicação de crise em situações de desastres que atenda às necessidades de pessoas surdas e pessoas cegas.

7. REFERÊNCIAS

[1] I. Aedo, S. Yu, P. Díaz, P. Acuña, and T. Onorati. Personalized alert notifications and evacuation routes in indoor environments. *Sensors*, 12(6):7804–7827, 2012.

[2] S. Banerjee, D. Mukherjee, and P. Misra. 'what affects me?': A smart public alert system based on stream reasoning. In *Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication*, ICUIMC '13, pages 22:1–22:10, New York, NY, USA, 2013. ACM.

[3] C. A. Bisol, C. B. Valentini, J. L. Simioni, and J. Zanchin. Deaf Students in higher education: Reflections on inclusion. *Cadernos de Pesquisa*, 40:147 – 172, 04 2010.

[4] D. o. H. A. DHA. Internationally agreed glossary of basic terms related to disaster management. Technical Report December 1992, 1992.

[5] D. P. Dhokane; and M. S. Wankhade. Survey of context-aware systems-challenges in development of context-aware applications. *International Journal of Application or Innovation in Engineering & Management*, 4(5):81–86, 2015.

[6] M. Klafft and H. G. Ziegler. A concept and prototype for the integration of multi-channel disaster alert systems. In *Proceedings of the 7th Euro American Conference on Telematics and Information Systems*, EATIS '14, pages 20:1–20:4, New York, NY, USA, 2014. ACM.

[7] A. Malizia, P. Acuna, T. Onorati, P. Diaz, and I. Aedo. CAP-ONES: an emergency notification system for all. *International Journal of Emergency Management*, 6(3/4):302, 2009.

[8] U. Meissen, M. Hardt, and A. Voisard. Towards a general system design for community-centered crisis and emergency warning systems. *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management*, pages 155–159, 2014. cited By 0.

[9] H. Mitchell, J. Johnson, and S. LaForce. The human side of regulation: Emergency alerts. In *Proceedings of the 8th International Conference on Advances in Mobile Computing and Multimedia*, MoMM '10, pages 180–187, New York, NY, USA, 2010. ACM.

[10] National Science and Technology Council (U.S.). Effective Disaster Warnings - Report by the Working Group on Natural Disaster Information Systems Subcommittee on Natural Disaster Reduction. Technical report, Subcommittee on Natural Disaster Reduction, Washington D.C., November 2000.

[11] OASIS. Common alerting protocol version 1.2. https://goo.gl/g1OJJG, 2010. Accessed in 21-05-2016.

[12] B. D. Phillips and B. H. Morrow. Social Science Research Needs: Focus on Vulnerable Populations, Forecasting, and Warnings. *Natural Hazards Review*, 8(3):61–68, 2007.

[13] H. T. Sullivan and M. T. Häkkinen. Preparedness and warning systems for populations with special needs: Ensuring everyone gets the message (and knows what to do). *Geotechnical and Geological Engineering*, 29(3):225–236, 2011.

[14] UNISDR. UNISDR Terminoology on Disaster Risk Reduction, 2009.

[15] United Nations. Global Survey of Early Warning Systems. Technical report, 2006.

[16] V. Vieira, P. Tedesco, and A. C. Salgado. Designing context-sensitive systems: An integrated approach. *Expert Systems with Applications*, 38(2):1119 – 1138, 2011. Intelligent Collaboration and Design.

[17] S. Wurster and U. Meissen. Towards an economic assessment approach for early warning systems: Improving cost-avoidance calculations with regard to private households. pages 439–443, 2014. cited By 1.

Detecção e Reconstrução de oclusões parciais em imagens de face visando Reconhecimento Biométrico

Alternative Title: Detection and Reconstruction of partial occlusions in face images for Biometric Recognition

Jonas M. Targino
Universidade de São Paulo
03828-000 São Paulo, Brasil
jonas.mendonca@usp.br

Clodoaldo A. M. Lima
Universidade de São Paulo
03828-000 São Paulo, Brasil
c.lima@usp.br

Sarajane M. Peres
Universidade de São Paulo
03828-000 São Paulo, Brasil
sarajane@usp.br

RESUMO

Há um crescente incentivo ao uso da tecnologia biométrica para melhorar, e até mesmo substituir, os métodos tradicionais de segurança. O campo da Biometria refere-se a uma grande variedade de tecnologias usadas para identificar ou verificar a identidade de uma pessoa por meio da mensuração e análise de vários aspectos fisiológicos e comportamentais do ser humano. Modalidades biométricas são características extraídas do corpo humano, que são únicas para cada indivíduo e que podem ser usadas para estabelecer sua identidade numa população. As principais modalidades biométricas empregadas são: impressão digital, face, voz, palma da mão, e íris. Dentre as modalidades biométricas, a face é a mais comumente vista e usada em nossa vida diária. Em aplicações do mundo real, o reconhecimento facial sofre de uma série de problemas nos cenários não controlados. Esses problemas são devidos, principalmente, a diferentes variações faciais que podem mudar muito a aparência da face, incluindo variações de expressão, de iluminação, alterações da pose, bem como oclusões parciais. Comparada com problemas de pose, iluminação e expressão, o problema relacionado à oclusão é relativamente pouco estudado na área. Há dois problemas distintos relacionados com o reconhecimento facial com oclusão: detecção da face ocluída e reconstrução da face ocluída. O objetivo desta pesquisa é investigar e avaliar técnicas para detecção e reconstrução de oclusões parciais em imagens de face, obtidos através de ambientes não cooperativos.

Palavras-Chave

Detecção, Oclusão, Face, Biométrico, Reconhecimento

ABSTRACT

There is a growing incentive to use biometric technology to improve, even replace, traditional security methods. The field of Biometrics refers to a wide variety of technologies used to identify or verify a person's identity by measuring and analyzing various physiological and behavioral aspects of the human being. Biometric modes are characteristics drawn from the human body, which are unique to each individual and can be used to establish their identity in a population. The main biometric modalities used are: fingerprint, face, voice, palm, and iris. Among the biometric modalities, the face is the most commonly seen and used in our daily life. In real-world applications, facial recognition suffers from a number of problems in uncontrolled scenarios. These problems are mainly due to different facial variants that can greatly change the appearance of the face, including variations in expression, lighting, changes in the pose, as well as partial occlusions. Compared with problems of pose, illumination and expression, the problem related to occlusion is relatively little studied in the area. There are two distinct problems related to facial recognition with occlusions: detection of the occluded face and reconstruction of the occluded face. The objective of this research is to investigate and evaluate techniques for detection and reconstruction of partial occlusions in face images obtained through non-cooperative environments.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•**Computing methodologies** → **Biometrics;** *Image segmentation; Reconstruction; Feature selection;*

Keywords

Detection, Occlusion, Face, Biometric, Recognition

1. INTRODUÇÃO

Segurança tem se tornado imprescindível na sociedade moderna devido ao grande volume de informações geradas diariamente. Mediante essa enorme quantidade de informações e exigências provindas do processo de globalização, as estratégias tradicionais de reconhecimento de identidade tais como números PIN, tokens, senhas e cartões de identificação tornaram-se obsoletas [18]. Além disso, essas técnicas são facilmente fraudadas, trazendo preocupações no que diz respeito à sua facilidade de aquisição e utilização por parte de terceiros não-autorizados. Neste contexto, existe um crescente incentivo ao uso da tecnologia biométrica para melhorar, e até mesmo substituir, os métodos tradicionais de

segurança. A Biometria refere-se a uma grande variedade de tecnologias usadas para identificar ou verificar a identidade de uma pessoa por meio da mensuração e análise de vários aspectos fisiológicos e comportamentais do ser humano [7]. Modalidades biométricas são características extraídas do corpo humano, que são únicas para cada indivíduo, podendo ser usadas para estabelecer sua identidade numa população. Existe uma quantidade considerável de modalidades biométricas, dentre as principais temos [9]: impressão digital [8], face [6], voz [2], palma da mão [12] e íris [20].

O processo de identificação biométrica pode ser dividido nas seguintes etapas: aquisição/segmentação, extração / seleção de características, comparação de características e armazenamento[18]. Nos últimos anos, o reconhecimento biométrico sofreu grandes avanços significativos em termos de confiabilidade e precisão [13], e algumas modalidades biométricas têm alcançado um bom desempenho em aplicações práticas. No entanto, mesmos os sistemas biométricos mais avançados ainda enfrentam alguns problemas [8].

Dentre as modalidades biométricas, a face é a mais comumente vista e usada em nossa rotina diariamente [4]. Desde o advento da fotografia, tanto órgãos governamentais e organizações privadas têm mantido banco de dados de fotografias de pessoas contendo a face (por exemplo, para identificação pessoal, passaportes, cartões de sócio). A face é uma característica universal, única para cada pessoa e possui boa aceitabilidade em ambientes de captura [19]. Atualmente, há progresso significativo em reconhecimento automático de face em condições controladas. Entretanto, a performance em condições não controladas é ainda insatisfatória. Sistemas de reconhecimento facial em ambientes de mundo real, frequentemente, lidam com condições não controladas e não previsíveis tais como grande mudança na iluminação, pose, expressão e oclusão, as quais introduzem variações intraclasse e degradam a performance de reconhecimento. Comparada com problemas de pose, iluminação e expressão, o problema relacionado à oclusão é relativamente pouco estudado na área. Os trabalhos de [1, 25] apresentam algumas dificuldades encontradas ao lidar com oclusões em imagens de face coletadas em um ambiente não controlado.

2. APRESENTAÇÃO DO PROBLEMA

A figura 2 apresenta três imagens de faces parcialmente ocluídas. Analisando esta figura é possível perceber que o óculos e o cachecol estão atuando como oclusões parciais da face, a qual impede a identificação automática por meio de técnicas tradicionais de reconhecimento facial. Uma ideia intuitiva para atacar oclusões no reconhecimento facial é detectar as regiões ocluídas e então realizar o reconhecimento usando somente as partes não ocluídas. No entanto, os tipos de oclusões são imprevisíveis em cenários práticos e podem impedir à identificação do indivíduo. A localização, tamanho e forma das oclusões são desconhecidas, aumentando assim a dificuldade de segmentação das regiões ocluídas das imagens de face. Atualmente, a maioria dos detectores de oclusão são treinados em faces com tipos específicos de oclusões e, portanto, generalizam mal para os diversos tipos de oclusões presentes em ambientes reais.

Embora tenha sido dada pouca atenção ao problema de oclusão na literatura de reconhecimento facial, a importância do mesmo deve ser enfatizada, pois a presença de oclusão é muito comum em cenários não controlados e pode estar associada a várias questões de segurança. Do ponto de vista do usuário, as oclusões faciais podem ocorrer por várias razões sejam elas intencionais ou não. Em primeiro lugar, acessórios faciais como óculos de sol, cachecol, maquiagem facial e chapéu / boné são bastante comuns na rotina diária de inúmeras pessoas. Algumas pessoas também usam véus por convicções religiosas ou hábitos culturais. Nos aplicativos de vigilância, a facilidade de uso é a propriedade mais importante que deve ser considerada, onde nenhuma cooperação do usuário pode ser esperada. O sistema deve ser capaz de reconhecer as pessoas, não importa quão grande seja esse ruído. Em segundo lugar, a oclusão também está aparecendo em cenários de segurança. Por exemplo, a máscara cirúrgica, cujo procedimento é necessário nas áreas restritas dos hospitais, e é frequentemente usada por pessoas na Ásia Oriental (por exemplo, China, Japão) para evitar a exposição à poluição do ar, doenças respiratórias ou alergia ao pólen. Também nas áreas de construção, capacete de segurança é vital para seres humanos por questões preventivas em tais áreas. Por último, mas não menos importante, as oclusões faciais são muitas vezes relacionadas a várias questões de segurança graves. Os hooligans do futebol inglês e os ladrões de caixas eletrônicos tendem a usar cachecóis e / ou óculos de sol para impedir que suas faces sejam reconhecidas. Ladrões de bancos e lojas geralmente usam um boné ao entrar em lugares onde cometem ações ilegais.

Figura 1: Imagens de faces parcialmente ocluídas

Atacar todas as oclusões mencionadas acima no reconhecimento facial é essencial para fins de segurança e aplicação da lei. Conforme mencionado anteriormente, identificar as pessoas sem qualquer cooperação na remoção de oclusão devido a acessórios faciais traz grande conveniência e confortabilidade para os usuários em inúmeros cenários. Por outro lado, identificar a presença de oclusões em locais restritos (por exemplo, hospital, área de construção) e revelar a identidade das pessoas nessas áreas garantem a segurança no ambiente. Da mesma forma, a detecção da presença de oclusão pode identificar pessoas suspeitas em certas áreas (por exemplo, estádio de futebol, caixa eletrônico, lojas, aeroportos) e o reconhecimento facial (apesar da presença de oclusão) nessas áreas pode ajudar a polícia a encontrar criminosos / fugitivos. Em suma, o reconhecimento de faces parcialmente ocluídas é muito importante e tem muitas aplicações potenciais em vigilância.

3. TRABALHOS RELACIONADOS

A metodologia clássica para tratar o reconhecimento biométrico baseado em imagens de face ocluídas consiste em encontrar características ou classificadores tolerantes à ruído. Vários trabalhos têm demonstrado que os algoritmos locais são menos sensíveis a oclusões parciais. Em [16] foi proposto Análise de Características Locais para extração de características usando estatísticas de segunda ordem. Martinez [14]

propôs uma abordagem que consiste em dividir a imagem da face ocluída em k regiões locais e para cada região é extraído os autovetores. Se uma região estiver ocluída, esta é automaticamente detectada usando um modelo probabilístico. Além disso, foi proposto uma ponderação para regiões locais a fim de fornecer robustez em problemas que envolvem variação de expressão. Tan *et al* [21] estendeu o trabalho de Matinez usando Mapas Auto-Organizáveis para apreender o subespaço ao invés de usar Gaussianas ou Mistura de Gaussianas. Em[11] foi proposto um método chamado Analises de Componentes Independente baseado em Saliências Locais, o qual emprega informação de saliências locais na extração das componentes independentes. Já em [5] foi proposto a combinação de métodos baseados em subespaço, como Análise de Componentes Principais, com métodos que visam a discriminação, como Análise de Discriminante Linear, objetivando melhor reconstrução das imagens de face. Em [10], foi proposto o emprego de Máquinas de Vetores Suporte Parcial em cenários onde a oclusão pode ocorrer em ambos os conjuntos de treinamento e teste. Neste caso, problema de oclusão foi tratado como um problema de reconstrução e a classificação foi realizado de acordo com o erro de reconstrução obtido para a imagem de teste.

Mais recentemente, a Classificação baseada em Representação Esparsa (SRC) tem alcançado desempenhos impressionantes no reconhecimento de imagens de face ocluídas. Wright *et al* [22] foi o primeiro a empregar SRC para reconhecimento de faces ocluídas. Neste trabalho, a face ocluída é representada como uma combinação linear de todas as imagens de face e um vetor de erros no nível de pixel. A classificação foi realizada por meio da minimização da norma *l1*. Yang *et al* [23] propuseram um método chamado Codificação Esparsa Robusta, que maximiza a estimativa da máxima verossimilhança do problema de codificação esparsa para oclusões não Gaussianas / Laplacianas de forma iterativa. Embora os métodos baseados em representação esparsa tenham obtido resultados de identificação significativos em faces ocluídas, esses métodos dependem de um grande número de amostras de faces de cada individuo com variações suficientes. Entretanto, em muitos cenários práticos de reconhecimento facial, as amostras de treinamento de cada indivíduo são muitas vezes insuficientes, no caso extremo somente uma face de cada individuo pode estar disponível.

Recentemente, alguns trabalhos revelaram que o conhecimento prévio de oclusões pode melhorar significativamente a precisão do reconhecimento de face baseado em informações locais. Rama et al [17] mostrou empiricamente que o conhecimento prévio sobre a oclusão (anotado manualmente) pode melhorar o desempenho do *Eigenface*. Zhang *et al* [24] propuseram usar a divergência de Kullback-Leibler para estimar a distribuição de probabilidade de oclusões no espaço de característica, de modo a melhorar o método LGBPHS (*Local Gabor Binary Pattern Histogram Sequence*) para a face parcialmente ocluída.

4. PROPOSTA DE SOLUÇÃO

Há várias técnicas propostas na literatura para detecção e reconstrução de oclusão em imagens de face. No entanto, não há nenhum estudo que mencione quais técnicas são mais adequadas para um determinado tipo de oclusão. Além disso, muitas dessas técnicas não foram propostas especificamente para reconhecimento biométrico. Neste trabalho pretende-se avaliar o impacto dessas técnicas na detecção e reconstrução de diferentes tipos de oclusão para o reconhecimento biométrico. Como resultado deste estudo, espera-se identificar os principais prós e contras dessas técnicas, como também eventuais limitações, e com isso sugerir qual a técnica mais adequada para determinado tipo de oclusão.

A presente proposta apresenta um grau significativo de contribuição em termos gerais, abrangendo o estado de arte, suas técnicas, algoritmos para detecção e reconstrução de oclusão e condições de viabilidade de aplicação, levando em conta as variações de iluminação, pose e oclusão. Os resultados obtidos com esta pesquisa serão de grande valia para pesquisadores que estão trabalhando na área, como também iniciantes que almejam conhecimento holístico do estado de arte para desenvolvimento de futuras pesquisas, de modo a produzir um conhecimento significativo e enriquecimento para a área de biometria e para os pesquisadores que almejam colaborar no reconhecimento facial independente das condições do ambiente de coleta de dados.

Em síntese, o estudo comparativo das técnicas de detecção de oclusões parciais e reconstrução de faces pode contribuir significativamente para a área de reconhecimento facial em ambientes não controlados, visto que a comparação envolve uma descrição detalhada de cada técnica e uma análise e síntese das suas semelhanças e diferenças. De acordo com [3] o estudo comparativo tem elevado grau de contribuição científico, dado que ele sugere similaridades e contrastes entre os casos, podendo participar da descoberta indutiva de novas hipóteses e posteriormente construir novas teorias.

5. AVALIAÇÃO DA SOLUÇÃO

Após a aplicação das técnicas de detecção e reconstrução, será realizada a extração de característica e posteriormente a classificação visando reconhecimento biométrico. As técnicas de detecção e reconstrução serão avaliadas em termos da taxa de reconhecimento alcançada usando diferentes classificadores (Redes Neurais Artificiais, Máquinas de Vetores Suporte e Floresta de Caminhos Ótimos). Teste de significância estatística será realizado para identificar as melhores e piores técnicas para determinado tipo de oclusão.

6. ATIVIDADES JÁ REALIZADAS

Como parte do desenvolvimento deste projeto, já foi realizado um estudo exploratório referente á biometria, com ênfase em reconhecimento biométrico baseado em imagens de face com oclusão parcial. Após este estudo, foi realizada uma revisão sistemática focada em estudos que tratam de detecção ou reconstrução de imagens de face com a presença de oclusões parciais. Com base nesta revisão, foi possível identificar que há um número considerável de técnicas para detecção e reconstrução de oclusões parciais em imagens de face. Entretanto, muitas dessas técnicas não conseguiram alcançar uma alta acurácia de identificação a ponto de serem aplicadas em situações cotidianas. Existem vários estudos que analisam a oclusão considerando outras modalidades biométricas tais como ouvidos, íris e nariz. A transferência de conhecimento entre domínios, pode ser de grande valia para o presente projeto. Por fim, já foram realizados alguns experimentos preliminares envolvendo processamento de imagens com o auxílio do software Matlab [1].

[1] https://www.mathworks.com/products/matlab.html

7. CONCLUSÃO

Este trabalho tem como objetivo principal investigar, implementar, avaliar e comparar as técnicas para detecção e reconstrução de oclusões parciais em imagens de face visando reconhecimento biométrico. Dentre as atividades exercidas, destaca-se um estudo exploratório e uma revisão sistemática da literatura abrangendo técnicas para detecção e reconstrução de oclusões parciais em imagens de face. Como diretrizes futuras, pretende-se utilizar a base de dados AR [15], disponível no link http://www2.ece.ohio-state.edu/˜aleix/ARdatabase.html, e implementar as principais técnicas de detecção e reconstrução de oclusões parciais em imagens de face encontradas na revisão sistemática realizada.

As atividades a serem desenvolvidas consistem em implementar as técnicas para detecção e reconstrução de oclusões parciais, analisar, comparar e distinguir suas especificidades de acordo com as variações de iluminação, pose e oclusão. A partir dessa comparação, propor quais técnicas se adequam melhor para tal contexto sobre tais variações. Após a implementação de cada técnica pretende-se descrever a mesma em detalhes de modo a construir um documento com notação unificada contendo as técnicas para detecção e/ou reconstrução de oclusões parciais em imagens de face presentes na literatura. Além disso, serão destacados os pontos fortes, fracos e eventuais limitações de cada técnica. A avaliação será realizada levando em conta a taxa de reconhecimento. Teste de significância estatística será realizado no sentido de identificar as melhores e piores técnicas para os diferentes tipos de oclusões parciais em imagens de face.

8. REFERÊNCIAS

[1] A. Aisha, S. Muhammad, S. J. Hussain, and R. Mudassar. Face recognition invariant to partial occlusions. *KSII Transactions on Internet and Information Systems (TIIS)*, 8(7):2496–2511, 2014.

[2] B. Beranek. Voice biometrics: success stories, success factors and what's next. *Biometric Technology Today*, 2013(7):9 – 11, 2013.

[3] D. Collier. The comparative method. In *Political Science: The State of the Discipline II*, pages 105–119. American Political Science Association, 1993.

[4] Y. Deng, Q. Dai, and Z. Zhang. Graph laplace for occluded face completion and recognition. *IEEE Trans. on Image Processing*, 20(8):2329–2338, 2011.

[5] S. Fidler, D. Skocaj, and A. Leonardis. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):337–350, March 2006.

[6] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *Proceedings of the IEEE Conf. on Comp. Vision and Pattern Recognition*, 2014.

[7] M. Hassaballah and S. Aly. Face recognition: challenges, achievements and future directions. *IET Computer Vision*, 9(4):614–626, 2015.

[8] A. K. Jain, P. Flynn, and A. A. Ross. *Handbook of Biometrics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[9] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. on circuits and systems for video tech.*, 14(1):4–20, 2004.

[10] H. Jia and A. M. Martinez. Support vector machines in face recognition with occlusions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–141, June 2009.

[11] J. Kim, J. Choi, J. Yi, and M. Turk. Effective representation using ica for face recognition robust to local distortion and partial occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1977–1981, Dec 2005.

[12] A. Kong, D. Zhang, and M. Kamel. A survey of palmprint recognition. *Pattern Recognition*, 42(7):1408 – 1418, 2009.

[13] M. A. Lone, S. Zakariya, and R. Ali. Automatic face recognition system by combining four individual algorithms. In *Computational Intelligence and Communication Networks (CICN), 2011 Int. Conf. on Communication Systems*, pages 222–226. IEEE, 2011.

[14] A. M. Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. on Pat. Analysis and Machine Intelligence*, 24(6):748–763, Jun 2002.

[15] A. M. Martinez and R. Benavente. The AR Face Database. Technical report, CVC, June 1998.

[16] P. S. Penev and J. J. Atick. Local feature analysis: a general statistical theory for object representation. *Network: Comp. in Neural Sys.*, pages 477–500, 1996.

[17] A. Rama, F. Tarres, L. Goldmann, and T. Sikora. More robust face recognition by considering occlusion information. In *8th IEEE Int. Conf. on Automatic Face Gesture Recognition*, pages 1–6, Sept 2008.

[18] M. Sharif, S. Mohsin, and M. Y. Javed. A survey: Face recognition techniques. *Research Journal of Applied Sciences, Engineering and Tech*, 4(23):4979–4990, 2012.

[19] J. Shermina and V. Vasudevan. Recognition of the face images with occlusion and expression. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(03), 2012.

[20] N. F. Soliman, E. Mohamed, F. Magdi, F. E. A. El-Samie, and A. M. Efficient iris localization and recognition. *Optik - International Journal for Light and Electron Optics*, 140:469 – 475, 2017.

[21] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang. Recognizing partially occluded, expression variant faces from single training image per person with som and soft k-nn ensemble. *IEEE Transactions on Neural Networks*, 16(4):875–886, July 2005.

[22] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(2):210–227, Feb 2009.

[23] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In *CVPR 2011*, pages 625–632, June 2011.

[24] B. Zhang, S. Shan, X. Chen, and W. Gao. Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition. *IEEE Trans. on Image Processing*, 16(1):57–68, Jan 2007.

[25] J. Zhu, D. Cao, S. Liu, Z. Lei, and S. Z. Li. Discriminant analysis with gabor phase for robust face recognition. In *Biometrics (ICB), 2012 5th IAPR International Conference on*, pages 13–18. IEEE, 2012.

Mediações de Conflitos no Poder Judiciário (MARCs): Alternativas tecnológicas para aproximação cidadã

Alternative Title: Mediation of Conflicts in the Judiciary (ADR) - Technological Alternatives for Citizen Approach

Emmanuel Pires
Núcleo de Pesquisa e Inovação em Ciberdemocracia
Programa de Pós-Graduação em Informática Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro – Brasil
emmanuel.pires@uniriotec.br

Renata Mendes de Araujo
Núcleo de Pesquisa e Inovação em Ciberdemocracia
Programa de Pós-Graduação em Informática Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro – Brasil
renata.araujo@uniriotec.br

RESUMO

A mediação de conflitos (MC) é um instrumento de pacificação dos conflitos sociais utilizado pelo Poder Judiciário (PJ). Além de promover o diálogo entre as partes conflitantes, a aproximação, a busca na solução do problema de forma duradoura, aprimora o conceito de política social de inclusão do cidadão de acesso à Justiça de forma gratuita. Contudo, um grande desafio que concerne ao PJ é disseminar para a sociedade a MC, pois o cidadão desconhece o seu processo de funcionamento, ou até mesmo, o fato da sua existência, tendo em vista que a sua regulamentação ocorreu somente em 2010. Esta pesquisa objetiva gerar um conjunto de recomendações e propostas de soluções inovadoras, tecnológicas ou não, para ampliar a capacidade de aproximação do cidadão das MCs.

Palavras-chave

Mediação de Conflitos, Prestação de Serviços Público, Colaboração, Ciberdemocracia.

ABSTRACT

Conflict mediation (MC) is an instrument for pacification of social conflicts used by the Judiciary (PJ). In addition to promoting dialogue between the conflicting parties, the approach, the search for solving the problem in a lasting way, improves the concept of social inclusion policy for citizens of access to justice for free. However, a major challenge for the PJ is to disseminate MC to society, since the citizen is unaware of its functioning, or even the fact of its existence, since its regulation occurred only in 2010. This research aims to generate a set of recommendations and proposals for innovative solutions, technological or not, to increase the capacity of approaching the citizen of MCs.

CCS Concepts

• Applied computing → Computers in other domains → Computing in government → E-government.

Keywords

Conflict Mediation, Public Service Delivery, Collaboration, Cyberdemocracy.

1. INTRODUÇÃO

A Democracia Eletrônica apoiada pela Tecnologia da Informação e Comunicação (TIC) é um dos instrumentos do Governo para melhorar a qualidade dos seus processos e da prestação de serviços públicos[10]. Aproximar o cidadão desses serviços torna-se essencial para que as discussões democráticas aconteçam. Entre os serviços prestados, encontra-se a mediação de conflitos (MC) e por ser considerada nova para o contexto do Direito Brasileiro, traz grandes desafios ao Conselho Nacional de Justiça (CNJ) para a sua consolidação como política social inclusiva de acesso à Justiça e como alternativa de pacificação de conflitos.

No entanto, o cidadão ainda desconhece os meios alternativos de resolução de conflitos (MARCs) ofertados pelo PJ, ora por não saber da sua existência, ora por desconhecer seu processo de funcionamento. Este fato se torna oneroso para o cidadão e para o PJ, porque ao se utilizar da via judicial se eleva ainda mais o número de estoque de processos pendentes de uma solução em todas as suas instâncias, os seus custos diretos e indiretos incidentes na tramitação [9] e, por conseguinte, gera o descrédito no PJ.

Portanto, esta pesquisa tem por abordagem inicial, analisar o contexto de funcionamento dos Centros Judiciários de Solução de Conflitos e Cidadania do RJ (CEJUSCs-RJ), mapear os seus processos e analisá-los à luz dos conceitos das disciplinas de *Business Process Management* (BPM) e Democracia Eletrônica. Por fim, gerar um conjunto de recomendações e soluções inovadoras, tecnológicas ou não para ampliar a capacidade de aproximação do cidadão dos CEJUSCs-RJ e estimulando-o a utilizar os MARCs ofertados pelo PJ.

Este trabalho está divido em 6 seções, incluindo esta introdução. A seção 2 traz a apresentação do problema. A seção 3 traz a proposta para uma possível solução. A seção 4 a avaliação da solução. Já a seção 5 as atividades já realizadas e 6, as conclusões.

2. APRESENTAÇÃO DO PROBLEMA

A MC já é utilizada em vários países há mais de 20 anos como uma forma alternativa de pacificação de conflitos [16]. No Brasil, a MC teve as suas origens em função de um cenário antigo que se

configurou na Justiça Brasileira, tendo a nossa Constituição Federal 1988 (CF88)[2], em seu artigo 5º, inciso XXXV promovido o direito fundamental de acesso à Justiça.

Outro fato importante, foi que após a CF88 algumas Leis e Resoluções contribuíram para fomentar e institucionalizar os meios de pacificação de conflitos. Em breves linhas, historicamente, destacam-se: a Lei nº 9.099/1995 [5] que institui em todas as unidades da federação os Juizados Especiais (Cível e Criminal) para litígios cujos valores não venham a ultrapassar a 40 (quarenta) salários mínimos; já a Resolução nº 125 [8] de 2010 do CNJ, traz um grande diferencial no cenário da MC, o qual normatiza, regulamenta e cria os Núcleo Permanente de Métodos Consensuais de Solução de Conflitos (NUPEMECs) e os CEJUSCs, em cada unidade da federação, com intuito de consolidar a política social de facilitação de acesso à Justiça para o cidadão. Somado a este fato, em 2015, entra em vigência a Lei Federal 13.140 [4], que dispõe sobre a MC e traz em seus artigos o entendimento dos conceitos e aplicações dos MARCs; e na sequência dos eventos, em 2016, entra em vigor o novo Código de Processo Civil (CPC) [3] de 2015, que além de reforçar, aprimorar e consolidar os MARCs traz em seu artigo nº 334, caput, a obrigatoriedade de utilização em todos os processos cíveis, cabíveis de audiências prévias de conciliação ou mediação.

Apesar do cenário promissor acima descrito, a MC ainda é desconhecida pela sociedade civil – aqui tratada como cidadão - como um meio de pacificação de conflitos. Como ilustração, a Figura 2 apresenta um panorama trazido pelo relatório anual Justiça em Número 2016 [7], ano base 2015, com os percentuais de índices de conciliações, i.e., um dos MARCs. Percebe-se que na Justiça Estadual, um dos itens do escopo desta pesquisa, um percentual de utilização de conciliações é de 9% e no PJ de 11%.

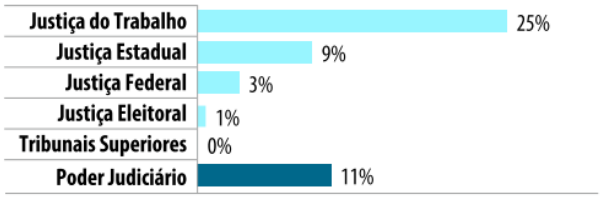


Figura 1. Índice de Conciliação do PJ, [7]

Portanto, esta foi a motivação devido ao fato de perceber que os MARCs são poucos utilizados por parte do cidadão. Identificar, esclarecer, apontar as causas raízes e as possíveis soluções tecnológicas ou não são alguns dos requisitos que fomentam esta pesquisa e o Grupo de Pesquisa e Inovação em Ciberdemocracia (CIBERDEM) do curso de Pós-Graduação da UNIRIO/PPGI para ampliar a capacidade de governança, transparência, participação social, educação e promoção da democracia.

3. PROPOSTA DE SOLUÇÃO

Com intuito de propor uma solução, foi realizado um levantamento bibliográfico de trabalhos correlatos em estudos que abordam critérios baseados em iniciativas para o engajamento e a participação do cidadão na relação governo-cidadão e que faz uso de TICs.

Entre estes estudos, destacam-se Araujo e Taher [1] cujos trabalhos propõem um *framework* como um facilitador na identificação dos requisitos necessários à participação do cidadão em diferentes níveis de contexto de prestação de serviços públicos. Estes requisitos servem para determinar quais ferramentas de TICs poderão dar suporte adequado na prestação de serviços. Com uma proposta de aumentar a interação entre os cidadão e o Governo, Diirr [10] discorre em seu trabalho a necessidade de capturar conversas sobre os processos de prestação de serviços públicos para melhoria da interação entre estes cidadãos e os prestadores de serviços públicos. Segundo Diirr [10], para uma organização que deseje implantar projetos de apoio à Democracia Eletrônica são necessários três pontos fundamentais: primeiro, **compreensão do contexto e seus processos** cujo objetivo é identificar como os processos são executados; segundo, a compreensão da **análise cultural** que objetiva entender os aspectos relacionados à cultura pré-existente da organização; e o terceiro, a **derivação de requisitos segundo os aspectos de apoio à participação democrática** a identificar as necessidades e requisitos necessários para endereçar os aspectos de colaboração, memória e transparência que apoiam à participação democrática.

Como referencial teórico, para dar o suporte necessário a esta pesquisa, serão utilizados abordagens de gestão de BPM e de conceitos de Democracia Eletrônica. Nos itens 3.1 a 3.3 estão apresentam-se resumidamente os respectivos conceitos das áreas que dão o alicerce a esta pesquisa.

3.1 BPM

Business Process Management (BPM) é um conjunto de princípios, métodos e ferramentas para projetar, analisar, executar e monitorar os processos de negócios [11]. Segundo [13] o BPM segue o conceito de uma disciplina que consiste na otimização, gerenciamento e execução do processo de negócio e obedece aos seguintes estágios: desenho dos processos/descoberta, configuração/implementação, aprovação/execução e avaliação.

3.2 Colaboração e o modelo ColabMM

A colaboração é um dos fatores determinantes para o sucesso e crescimento de qualquer negócio, não importando ele ser empresarial ou pessoal. O termo colaboração é definido segundo o dicionário digital do Aurélio como: "trabalhar em comum com outrem; agir com outrem para a obtenção de determinado resultado".

O Modelo de Maturidade de Colaboração (ColabMM) define os níveis de maturidade que uma organização pode atingir no que tange à colaboração em seus processos de negócio [15]. O objetivo é organizar as práticas de colaboração que podem ser aplicadas à modelagem dos processos de negócio e apoiar as organizações a inserirem e estimularem a colaboração nos processos de negócio. São definidos quatro níveis colaboração para o ColabMM: Casual, Planejado, Perceptivo e Reflexivo. No **Casual** a colaboração não está explícita; No **Planejado**, os processos começam a ser modificados e há inclusão de atividades de colaboração com planejamento; No **Perceptivo**, os membros do grupo já conhecem as responsabilidades e sabem quais tarefas a executar e por fim, o no **Reflexivo** onde as organizações percebem o valor do conhecimento que está sendo gerado [15][18].

3.3 Transparência e Memória

Além do habilitador de **colaboração**[15][12], temos os de **transparência**[6] e **memória**[12] como importantes habilitadores específicos de Democracia Eletrônica. Tornar os serviços públicos transparentes se faz necessário para permitir uma melhor visão sobre os processos de funcionamento, informações e conhecimento dos serviços prestados e aumentar o grau de confiança entre as organizações e a sociedade civil[6].

Outra questão importante para Democracia Eletrônica é a questão do registro das informações que ocorrem durante as discussões e

deliberações de assuntos públicos. Muitas das vezes essas discussões entre o cidadão e Governo para melhorias dos processos e serviços prestados, críticas, sugestões, documentos formais gerados nas organizações, que se traduzem no conhecimento tácito ou explicito, não são preservados. Portanto, organizar, armazenar, recuperar e compartilhar estas informações se faz necessário, pois auxiliam na melhoria das discussões que envolvem a prestação de serviços públicos[10].

Baseados nos estudos apresentados acima e através da análise do cenário dos processos dos CEJUSCs-RJ, onde ocorrem as mediações de conflitos, assim como, apoiado pelas áreas de conhecimento de BPM e Democracia Digital, será possível propor um **método para aproximação cidadã (MAC)**. Este método se propõe a descrever os passos necessários para que um especialista em processos (analista de processos ou de negócio) ao aplicá-lo, consiga fazer a análise deste processo e por conseguinte, o modifique e o torne o mais próximo do cidadão.

4. AVALIAÇÃO DA SOLUÇÃO

A metodologia a ser utilizada nesta pesquisa para avaliar o método proposto, será inicialmente um Estudo de Caso. Esta metodologia é a mais popular e bem estabelecido método científico. Agrega o conceito de pesquisa de um fenômeno em um ambiente natural onde são definidas as fronteiras entre o conhecido e o desconhecido[19].

Com intuito de avaliar a solução proposta, serão utilizadas as seguintes abordagens neste trabalho de pesquisa:

Primeira: ao término da descrição das atividades e das fases que compõem o MAC, será realizada, por um especialista em processos, uma "calibragem" para identificar as possíveis falhas quanto a sua eficácia e eficiência para futuro ajustes. **Segunda**: como forma de validação do método proposto, será realizada uma pesquisa qualitativa com especialistas em processos, tanto da indústria quanto do meio acadêmico, com o foco na viabilidade de aplicação, na execução e nos resultados esperados do artefato (MAC). **Terceira**: será analisar dados coletados, durante o Estudo de Caso, conforme definido em protocolo de execução, para posterior divulgação dos resultados ao meio científico.

Assim, através da análise dos dados coletados e a divulgação à comunidade científica, será possível dar credibilidade ao método concebido, para posterior aplicação em outros CEJUSCs ou em qualquer órgão público.

5. ATIVIDADES REALIZADAS

Para o método de pesquisa proposto, i.e., Estudo de Caso, até a presente data, foram realizadas entrevistas iniciais com os principais atores que trabalham na unidade de negócio do CEJUSC-RJ (Capital), entre: coordenador, auxiliares administrativos, equipe de apoio, estagiários e mediadores, para levantamento das rotinas e atividades.

Além disso, foram utilizados os conceitos e técnicas de mapeamento de processos e a ferramenta BPMN Bizagi[1] para as modelagens dos processos que compõem a **mediação pré-processual** (quando a mediação é praticada antes de se constituir um processo judicial) e a **mediação endoprocessual** (quando a mediação é praticada no transcorrer do processo judicial sendo esta ordenada por um Juiz de Direito).

O método proposto encontra-se na fase de concepção, com as definições das fases e das atividades necessárias para a sua aplicação. Os habilitadores usuais de BPM proposto por Sharp [20] (Recursos Humanos, Workflow, TI, Regras Políticas, Motivação e Métricas e Infraestrutura) e os específicos de Democracia Eletrônica (Colaboração, Transparência e Memória) [10][12][15], estão sendo avaliados para se saber quais farão parte do método, ou seja, quais habilitadores farão parte do método para análise de um processo de aproximação cidadã.

Em paralelo, está sendo confeccionado um "Book de Processos" que conterá o detalhamento dos processos do CEJUSC-RJ das **mediações pré-processual e endoprocessual**. O Book será entregue ao NUPEMEC-RJ ao final dos trabalhos desta pesquisa, com as respectivas sugestões de melhorias dos seus processos para a aproximação do cidadão.

6. CONCLUSÃO

A proposta apresentada neste trabalho consiste em analisar e utilizar os processos iniciais (AS-IS) levantados do CEJUSC-RJ e as áreas de conhecimento de BPM e de Democracia Digital a fim de analisar e identificar os potenciais problemas que levam o cidadão a não utilizar na plenitude os MARCs ofertados pelo PJ e em seguida, propor um método para aproximação cidadã.

Como consequência, ao final deste trabalho de pesquisa, são esperados que as seguintes contribuições sejam alcançadas: **i)** para a área de BPM, um método como um facilitador para análise e melhoria de um processo público no quesito aproximação cidadã; **ii)** para a área de Democracia Eletrônica, contribuir e reforçar à utilização dos conceitos que envolvem os habilitadores de Democracia (colaboração, transparência e memória) na aplicação do método; **iii)** e para ambas áreas, que o método proposto (MAC) possa ser replicado em outros órgãos públicos.

Assim, adotando-se as abordagens das áreas de conhecimento mencionadas acima, espera-se gerar propostas de soluções inovadoras para potencializar a aproximação do cidadão dos respectivos CEJUSCs-RJ, ampliando os conceitos de empoderamento[17][9][14] e inclusão social do cidadão para acesso à justiça gratuita almejados tanto pelo NUPEMEC-RJ, quanto pelo CNJ.

7. REFERENCES

[1] Araujo, R. and Taher, Y. 2014. Refining IT Requirements for Government-Citizen Co-participation Support in Public Service Design and Delivery. *Conference for E-Democracy and Open Governemen*. May (2014), 61.

[2] Brasil. Constituição da República Federativa do Brasil: 1988. *http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm*. Accessed: 2016-12-10.

[3] Brasil. Lei Nº 13.105: 2015. *https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm*. Accessed: 2016-10-03.

[4] Brasil. Lei Nº 13.140: 2015. *http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2015/Lei/L13140.htm*. Accessed: 2016-05-22.

[5] Brasil. Lei Nº 9.099: 1995. *http://www.planalto.gov.br/ccivil_03/leis/L9099.htm*. Accessed: 2016-11-18.

[6] Cappelli, C. 2009. *Uma abordagem para transparência em processos organizacionais utilizando aspectos*. Tese (Doutorado em Informática). Pontifícia Universidade Católica do Rio de Janeiro-PUC.

[1] http://www.bizagi.com

[7] CNJ. Justiça em números 2016: ano-base 2015: 2016. *http://cnj.jus.br/programas-e-acoes/pj-justica-em-numeros*.

[8] CNJ. Resolução nº 125: 2010. *http://www.cnj.jus.br/busca-atos-adm?documento=2579*. Accessed: 2016-11-09.

[9] Cruz, S.G. da and Silva, F.A.N. da 2015. Conciliação, Mediação e Arbitragem. *Revista Ciências Jurídicas e Sociais*.

[10] Diirr, B. 2011. *Conversas sobre processos de prestação de serviços públicos*. Dissertação (Mestrado em Informática). Universidade do Estado do Rio de Janeiro-UNIRIO.

[11] Dumas, M. et al. 2013. *Fundamentals of Business Process Management*. Springer Berlin Heidelberg.

[12] Engiel, P. 2009. *Habilitando processos de prestação de serviços à particiapção e à democracia - o Caso da Escola de Informática Aplicada/UNIRIO*. TCC (Bacharel em Sistemas de Informação). Universidade do Estado do Rio de Janeiro-UNIRIO.

[13] Filipowska, A. et al. 2009. Procedure and guidelines for evaluation of BPM methodologies. *Business Process Management Journal*. 15, 3 (Jun. 2009), 336–357.

[14] Jurídico, N. and Araújo, M.D.E. 2015. Movimento em prol da adoção de soluções alternativas de conflitos e a nova lei de mediação. *Boletim Conteúdo Jurídico*. 480, ano VII (2015), 11–30.

[15] Magdaleno, A. and Araujo, R. 2009. A maturity model to promote collaboration in business processes. *Int. J. Business Process Integration and Management*. X, (2009).

[16] Maia, R.C.V.M. and Barbosa, V.P.O. 2012. Acesso à Justiça e formas de resolução de conflitos no Brasil e no mundo. *Conferência Internacional - Forum Munidal de Mediación, VIII.* (Valencia, Espanha, 2012), 183–189.

[17] Manual de Mediação Judicial-2015: 2015. *http://www.cnj.jus.br/noticias/cnj/79758-quinta-edicao-do-manual-de-mediacao-e-disponibilizada-pelo-cnj*. Accessed: 2016-05-30.

[18] Pimentel, M. et al. 2008. Um processo de desenvolvimento de sistemas colaborativos baseado no Modelo 3C: RUP-3C-Groupware. *Anais do IV SBSI. SBSI*. (2008), 35–47.

[19] Recker, J. 2013. *Scientific Research in Information Systems - A Beginner's Guide*. Springer Heidelberg New York.

[20] Sharp, A. and Mcdermott, P. 2008. *Workflow Modeling: Tools for Process Improvement and Application Development, Second Edition*.

Um Modelo de Gestão de Produto de Software para as Universidades Federais Brasileiras

Alternative Title: A Software Product Management Model for the Brazilian Federal Universities

Ana Klyssia Martins Vasconcelos
Universidade de São Paulo
Rua Arlindo Béttio, 1000, Ermelino Matarazzo
São Paulo, Brasil
ana.klyssia@usp.br

Marcelo Medeiros Eler
Universidade de São Paulo
Rua Arlindo Béttio, 1000, Ermelino Matarazzo
São Paulo, Brasil
marceloeler@usp.br

RESUMO

Governo Eletrônico é uma iniciativa que possibilita aos estados oferecerem produtos e serviços a outros governos, organizações públicas e privadas, turistas e principalmente cidadãos. As soluções de software fornecidas pelas universidades públicas federais brasileiras à comunidade acadêmica e à população também se enquadram nesta categoria, portanto devem seguir as leis, decretos e padrões definidos pelo governo federal. Uma pesquisa do TCU revelou que a maioria das instituições federais não atendem as expectativas do governo em relação aos métodos de desenvolvimento e à aderência aos padrões de e-government definido, principalmente por não ter um modelo de gestão apropriado. A proposta deste trabalho de mestrado é adaptar um modelo de Gestão de Produtos de Software consolidado na literatura para auxiliar instituições públicas federais, em especial as universidades federais, a gerir seus produtos de software de forma a atender às expectativas dos clientes, da instituição e das instâncias superiores do governo.

Palavras-Chave

gestão de produto de software, engenharia de software, ciclo de vida do produto, governo eletrônico, setor público

ABSTRACT

Electronic Government is an initiative that enables countries to offer products and services to other governments, public and private companies, tourists and mainly citizens. As software solutions for Brazilian federal public universities for the academic community and for society also fall into this category, following decrees and standards by the federal government. A TCU survey revealed that most federal institutions do not serve as expectations about government in development methods and adherence to e-government standards. A proposal of this master's work is a Software Products Management model consolidated in the literature for federal higher education institutions, especially as federal universities manage its software products in a way that meets the expectations of the clients, institution and Upper instances of government.

Categories and Subject Descriptors

K.6.3 [**Management of Computing and Information Systems**]: Software management—*software process* ; D.2.9 [**Software Engineering**]: Management—*lifecycle, software process models*

General Terms

Management

Keywords

Software product management, software engineering, product lifecycle, e-government, public sector

1. INTRODUÇÃO

Governo eletrônico pode ser definido, de forma geral, como uma iniciativa que possibilita aos Estados oferecerem informações, produtos e serviços a outros governos, organizações públicas e privadas, turistas e principalmente cidadãos por meio do uso de tecnologia de informação e comunicação [13]. As primeiras ações do Governo Eletrônico brasileiro foram registradas em 2000. Desde então, o governo federal criou comissões especiais, decretos, leis e padrões para definir as estratégias e os níveis de qualidade esperados dos serviços eletrônicos fornecidos por todas as instituições federais.

Dentre os padrões criados pelo governo federal para as soluções de governo eletrônico estão, por exemplo, o Modelo de Acessibilidade de Governo Eletrônico (e-MAG) [11], o Padrão de Interoperabilidade de Governo Eletrônico (e-PING) [12], e a Identidade Digital de Governo Eletrônico (IDG). Apesar de ter a 9ª maior economia do mundo, o Brasil ainda está na 51ª posição no ranking da ONU que avalia os países de acordo com a maturidade de seus sistemas de governo eletrônico, e 37º no ranking de participação [13].

As universidades federais brasileiras, assim como todas as instituições federais, também estão sujeitas às leis, decretos e padrões definidos pelo governo federal. Seus produtos

e serviços eletrônicos à população e à comunidade universitária enquadram-se na categoria de governo eletrônico, e portanto devem atender não só aos requisitos definidos para suas soluções, mas também as expectativas do governo federal.

2. APRESENTAÇÃO DO PROBLEMA

O Tribunal de Contas da União (TCU) realizou, recentemente, uma extensiva pesquisa nos setores de TI das instituições federais e identificou dois problemas principais: i) as instituições não entregam soluções que satisfazem os padrões definidos pelo governo federal [8]; ii) os recursos humanos são escassos para atender a alta demanda por soluções de software [3, 4]. Na maioria dos setores os métodos de desenvolvimento não são padronizados, muitos departamentos não seguem ou até mesmo desconhecem os padrões e estratégias de e-gov definidos pelo governo federal, e por isso há soluções que não atingem os objetivos esperados. Um exemplo recente de uma solução do governo que não atendeu as expectativas é o e-social[1].

Esses mesmos problemas são recorrentes no setor de educação e, em partiular, nos setores de TI das universidades federais. Em um estudo recente realizado na Universidade Federal de Uberlândia, constatou-se que a divisão de desenvolvimento de portais web criou e mantem mais de 600 portais web institucionais. Além da alta demanda por novos portais, a volatilidade dos requisitos do setor e as mudanças nas leis, decretos e padrões do governo federal obrigam as equipes a realizarem constantes manutenções nos portais existentes[6, 5].

Um dos grandes problemas neste contexto é a falta de um modelo estratégico de gestão que considere produtos de software padrões, ou linhas de produto, para reduzir o esforço de desenvolvimento e manutenção. Apesar de muitos portais serem muito semelhantes, cada um deles é desenvolvido como um projeto único, sob demanda, sem considerar o reúso de soluções anteriores em nenhum nível (requisitos, projeto, arquitetura, código, etc) [8]. Por exemplo, os portais das pró-reitorias da instituição são muitos semelhantes entre si e poderiam ser construídos como especializações de um portal padrão para as pró-reitorias. Em pesquisa recente, descobriu-se que outras universidades federais também enfrentam problemas semelhantes. Não é surpresa, portanto, que as poucas e pequenas equipes de desenvolvimento não consigam desenvolver soluções que atendam tantos às expectativas da instituição quanto às do governo federal[6, 5].

3. PROPOSTA DE SOLUÇÃO

A proposta de solução é que as instituições federais, e em particular as universidades federais brasileiras, adotem a estratégia de desenvolver soluções padronizadas ao invés de produtos personalizados. Para isso, propõem-se adoção de um modelo de Gestão de Produto de Software (GPS), cujo objetivo é reger um produto padrão desde sua criação para o mercado ou a entrega ao cliente e serviço, a fim de gerar o maior valor possível para uma empresa [14]. A disciplina de GPS tem o potencial no atendimento das exigências de fornecer produtos e serviços com requisitos voláteis e que exigem constante manutenção, alta qualidade técnica e ciclo de desenvolvimento rápido [10].

Essa disciplina tem se mostrado de grande importância no setor privado quanto à obtenção do sucesso do produto e consequentemente da organização, pois exerce influência direta na qualidade, rentabilidade e previsibilidade do produto. As práticas da GPS refletem-se em várias áreas, tais como gestão de portfólio, mapeamento de características comuns e específicas dos produtos, gestão de requisitos, definição de estratégia de marketing e desenvolvimento de produtos. Portanto, o objetivo deste projeto de mestrado é propor um modelo de GPS adaptado ao contexto de desenvolvimento de software das universidades públicas federais brasileiras.

Como os modelos de GPS foram criados principalmente para o setor privado, é necessário fazer adaptações dos modelos existentes para atender as características das instituições públicas. Para atingir este objetivo, as seguintes atividades foram definidas: i) Levantamento bibliográfico para identificar os modelos de GPS existentes; ii) mapeamento dos principais atores e atividades de um modelo de GPS no contexto de uma universidade pública federal; iii) aplicação de um survey nas universidades públicas federais brasileiras para identificar as principais semelhanças e diferenças na estrutura organizacional e no processo de desenvolvimento de software; iv) adaptação de um modelo de GPS para o contexto e realidade das universidades públicas federais brasileiras.

4. PROJETO DE AVALIAÇÃO DA SOLUÇÃO

O modelo de GPS proposto neste projeto de mestrado será avaliado pelo diretor e equipe de desenvolvimento de portais web da Universidade Federal X em um workshop a ser realizado nesta instituição. Além disso, o modelo será avaliado pela pesquisadora Inge van de Weerd, especialista em GPS e criadora do *Software Product Management Competence Model (SPMCM)* [2], com a qual estabelecemos uma colaboração para validar os resultados intermediários de nossa pesquisa. Entende-se que, idealmente, a validação do modelo de GPS criado para o contexto das universidades públicas federais deveria ser avaliado pela aplicação do modelo em diferentes universidades federais, seguida de uma avaliação qualitativa e quantitativa dos resultados alcançados após sucessivos ciclos de desenvolvimento. Entretanto, este tipo de validação não será possível em razão do curto período de um mestrado acadêmico.

5. ATIVIDADES JÁ REALIZADAS

Até o momento, as seguintes atividades foram realizadas: levantamento bibliográfico, mapeamento das características de GPS em uma universidade pública federal, e criação de um survey a ser submetido para as universidades federais brasileiras. No levantamento bibliográfico foram considerados os modelos de GPS propostos, incluindo um framework definido pelo SPMBOK (Software Product Management Body of Knowledge) [9]. Dentre os modelos analisados, optou-se por usar como referência o framework *Software Product Management Competence Model (SPMCM)* ou Modelo de Competência de Gerenciamento de Produto de Software proposto por Van de Weerd [2].

Este modelo foi escolhido por ser simples, apresentar os principais conceitos de GPS e mostrar o fluxo de dados en-

[1]O eSocial é um projeto do governo federal que unifica o envio de informações pelo empregador em relação aos seus empregados. Acesse: http://www.esocial.gov.br

Tabela 1: Resultado do mapeamento inicial

Áreas e atores	Modelo	Univ. Federal
Gestão de portfólio	Existente	Existente
Mapeamento de produto	Existente	Inexistente
Gestão de requisitos	Existente	Existente
Planejamento de versões	Existente	Existente
Governo Federal	Inexistente	Existente
Mercado	Existente	Existente
Parceiros	Existente	Existente
Clientes	Existente	Existente
Diretoria	Existente	Existente
Pesquisa e inovação	Existente	Inexistente
Serviços	Existente	Existente
Desenvolvimento	Existente	Existente
Suporte	Existente	Existente
Vendas	Existente	Inexistente
Marketing	Existente	Existente

tre os atores e atividades definidas pelo modelo, o mesmo é muito utilizado no meio acadêmico e existem trabalhos científicos relatando sua aplicação em empresas [1] e até em instituições de saúde [7], consolidando assim, a sua validade e importância. O modelo possui as quatro áreas principais de gestão de produto de software, a saber: gestão de requisito, gestão de portfólio, mapeamento do produto e planejamento de versões; atores externos(mercado, cliente e parceiros) e atores internos(diretoria, vendas, marketing, pesquisa e inovação, Desenvolvimento, suporte e serviços) [15].

Um mapeamento dos principais componentes deste modelo foi realizado na Universidade Federal X. O método utilizado para coleta de dados foi a realização de três entrevistas semi-estruturadas. Todas as entrevistas foram gravadas, transcritas e sumarizadas em um relatório de acordo com o roteiro elaborado para a entrevista. Os funcionários entrevistados tem experiência mínima de 6 anos dentro da instituição. A primeira entrevista durou duas horas e meia, pois foi preciso conhecer o vocabulário e as principais características deste setor. As duas outras entrevistas duraram em média 50 minutos.

Um resumo das diferenças e semelhanças encontradas entre o modelo de referência utilizado e o contexto de desenvolvimento da Universidade Federal X é apresentado na Tabela 1. Aqui são apresentadas somente as diferenças em termos de existente e inexistente, mas em alguns casos há diferenças na aplicação de cada atividade e o papel desempenhado pelos stakeholders, mas não há diferenças significativas que impeçam sua aplicação conforme proposto no modelo.

O resultado final dessa parte da pesquisa foi a reestruturação do modelo de competência de gestão de produto de software sob a visão da instituição de ensino federal estudado. Deste modo, ao final dessa etapa, o mapeamento foi validado pela especialista Inge Van de Weerd, autora do modelo de referência utilizado neste trabalho.

Após esta estapa, foi elaborado um survey que será enviado para as universidades públicas federais afim de evidenciar se as atividades/processos previstos no modelo adaptado também podem ser aplicados no contexto de outras universidades federais. O resultado do survey será importante para atualizar o mapeamento realizado e fornecerá informações para a identificação dos fluxos de dados e controle nas áreas de GPS do modelo considerando os atores identificados.

6. CONCLUSÃO

O setor público, como qualquer organização, precisa entregar soluções de software que atendam tanto os clientes quanto as estratégias organizacionais definidas pela instituição ou pelas instâncias superiores do governo. Atender à alta demanda por novas soluções e realizar manutenção das existentes é uma tarefa desafiadora para as reduzidas equipes de TI das instituições federais, inclusive das universidades federais.

Portanto, acredita-se que o desenvolvimento de produtos padronizados de software aliado à utilização de um modelo de GPS adaptado para a realidade e contexto das universidades federais será útil para a entrega de soluções não só com qualidade técnica, mas também alinhadas às estratégias de TI e organizacionais definidas pela instituição. Dadas as características semelhantes entre diversas instituições públicas, acredita-se que o modelo proposto é o passo inicial para estender o modelo para outras instituições fora do setor da educação.

Até o momento, o mapeamento do modelo de referência de GPS na Universidade Federal X revelou que há diferenças entre o que é previsto pelo modelo e a realidade da instituição analisada. Entretanto, nenhuma característica do setor público identificado é fator limitante para a aplicação de um modelo de GPS adaptado para os atores e as atividades possíveis neste contexto.

Espera-se que este trabalho contribua para as reflexões e para as ações que procuram melhorar os serviços digitais desenvolvidos pelo governo brasileiro, para que ele possa se aproximar cada vez mais das expectativas que todos temos da presença do governo no meio digital. Trata-se de serviços de qualidade, de novos canais de participação, de transparência e, acima de tudo, de uma capacidade de se adaptar e responder rapidamente as inovações tecnológicas e aos novos canais de comunicação que evoluem constantemente.

7. REFERÊNCIAS

[1] W. Bekkers, M. Spruit, I. van de Weerd, R. van Vliet, and A. Mahieu. A situational assessment method for software product management. In *ECIS 2010 PROCEEDINGS*, page 22. Educational Collaborative for International Schools, 2010.

[2] W. Bekkers, I. van de Weerd, M. Spruit, and S. Brinkkemper. A framework for process improvement in software product management. In *European Conference on Software Process Improvement*, pages 1–12. Springer, 2010.

[3] T. de Contas da União. Levantamento de governança de tecnologia da informação. `http://portal.tcu.gov.br/comunidades/fiscalizacao-de-tecnologia-da-informacao/atuacao/perfil-de-governanca-de-ti`, 2014.

[4] T. de Contas da União. Levantamento do pessoal de tecnologia da informação. `http://portal.tcu.gov.br/lumis/portal/file/fileDownload.jsp?fileId=8A8182A25232C6DE0152A26699517504`, 2015.

[5] A. D. de Oliveira and M. M. Eler. Interoperability in e-government solutions: The case of brazilian federal universities. In *In Proceedings of the Brazilian*

Symposium on Information Systems (SBSI) - to appear., pages 1–8. ACM, 2017.

[6] A. D. de Oliveira and M. M. Eler. Strategies and challenges on the accessibility and interoperability of e-government web portals: A case study on brazilian federal universities. In *In Proceedings of the IEEE Computers, Software, and Applications Conference (COMPSAC 2017) - to appear.*, pages 1–6. IEEE, 2017.

[7] S. A. Fricker, M. Persson, and M. Larsson. Tailoring the software product management framework for use in a healthcare organization: Case study. In *European Conference on Software Process Improvement*, pages 155–166. Springer, 2013.

[8] A. P. Garcia. Uma investigação sobre as dificuldades de planejamento de ti em instituições públicas brasileiras: Uma abordagem com teoria fundamentada em dados. Master's thesis, Universidade Federal de Pernambuco, 2016.

[9] ISPMA. *Software Product Management Body of Knowledge (SPMBoK)*. International Software Product Management Association, 2016.

[10] T. Kilpi. Product management challenge to software change process: preliminary results from three smes experiment. *Software Process: Improvement and Practice*, 3(3):165–175, 1997.

[11] S. d. L. e. T. d. I. Ministério do Planejamento, Orçamento e Gestão. Modelo de acessibilidade em governo eletrônico - e-mag. `http://emag.governoeletronico.gov.br`, 2014.

[12] S. d. L. e. T. d. I. Ministério do Planejamento, Orçamento e Gestão. Padrões de interoperabilidade de governo eletrônico - e-ping. `http://eping.governoeletronico.gov.br`, 2017.

[13] U. Nations. E-government for the future we want. *United Nations E-Government Survey 2014*, 2014.

[14] I. Van De Weerd, S. Brinkkemper, R. Nieuwenhuis, J. Versendaal, and L. Bijlsma. On the creation of a reference framework for software product management: Validation and tool support. In *Software Product Management, 2006. IWSPM'06. International Workshop on*, pages 3–12. IEEE, 2006.

[15] I. Van De Weerd, S. Brinkkemper, R. Nieuwenhuis, J. Versendaal, and L. Bijlsma. Towards a reference framework for software product management. In *Requirements Engineering, 14th IEEE International Conference*, pages 319–322. IEEE, 2006.

Analise de riscos em projetos de implementação de ERP influenciados por incertezas sazonais

Alternative Title: Risk assessment in ERP implementation projects influenced by seasonal uncertainties

Edmir Parada Vasques Prado
Universidade de São Paulo - EACH
Av. Arlindo Bettio, 1000
CEP: 03828-000 São Paulo/SP
055 11 3091-8893
eprado@usp.br

Paulo Mannini
Universidade de São Paulo - EACH
Av. Arlindo Bettio, 1000
CEP: 03828-000 São Paulo/SP
055 11 99437-0610
paulo.mannini@usp.br

RESUMO

O gerenciamento dos riscos constitui um dos pontos fundamentais para o sucesso de projetos de implementação de ERP. Um aspecto que influencia significativamente os projetos e que deveria ser considerado na análise de riscos é a sazonalidade, apesar de ser pouco abordado na literatura. Neste sentido, este trabalho trata-se de uma dissertação de mestrado em andamento e que tem como objetivo geral realizar uma pesquisa qualitativa, baseada na técnica Delphi, para definir quais são os métodos de análise de riscos mais adequados para se analisar riscos em projetos de implementação de sistemas *Enterprise Resource Planning* (ERP) influenciados por incertezas sazonais.

Palavras-chave

Gerenciamento de risco, Projeto de implantação de ERP, Sazonalidade.

ABSTRACT

The risk management constitutes one of the fundamental points for the success of ERP implementation projects. An aspect that influences significantly the projects and should be considered into risk analysis is the seasonality, although this has been not discussed in the literature. In this respect, this work is a master dissertation in progress that has as its general objective to perform a qualitative research, based on Delphi technique, to define which are the risk analysis methods more appropriated to analyze risks in ERP implementation projects influenced by seasonal uncertainties.

CCS Concepts

• **K.6.0 [MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS]:** General—*Economics* • **K.6.1 [MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS]:** Project and People Management techniques—*Management technique*.

Keywords

Risk Management, ERP implementation project, seasonality.

1. INTRODUÇÃO

Muitas organizações em diversos países têm implementado sistemas ERP desde os anos 90, para obterem uniformidade das informações entre seus sistemas de informação e reformularem seus negócios [14]. Porém, segundo *Standish Group* [17], baseado nos resultados de grandes projetos de software obtidos a partir do *CHAOS Report* entre os anos de 2003 e 2012, apenas 6% dos projetos são finalizados com prazo, custo e implementação satisfatória. Esta análise mostrou também que 52% dos projetos foram considerados desafiadores, por terem extrapolado o orçamento, ultrapassado prazos planejados e/ou não obterem uma implementação satisfatória. Um destaque para esta análise é que, infelizmente, os projetos que representam os 42% restantes da análise foram considerados como falhos, pois foram cancelados ou finalizaram sem a utilização do sistema após implementação.

De acordo com Tsai et al. [18], deficiências do planejamento e controle de riscos na implementação de ERP contribuem para as altas taxas de falha destes projetos, enquanto que um melhor entendimento do planejamento e controle dos fatores de risco podem ajudar aos gerentes de projeto a focar em áreas com alto potencial de riscos.

Considerando a grande quantidade de riscos envolvidos com os projetos de implementação de ERP e a influência dos riscos no sucesso destes projetos, *Project Management Institute* (PMI) [6] sugere que ”é preciso fazer uma escolha consciente em todos os níveis da organização para identificar ativamente e buscar o gerenciamento eficaz dos riscos”. Segundo o PMI [6] o gerenciamento de riscos é realizado por meio de seis processos, que tem como objetivos ”aumentar a probabilidade e o impacto dos eventos positivos e reduzir a probabilidade e o impacto dos eventos negativos no projeto”. Como consequência, a análise qualitativa e análise quantitativa de riscos, que são processos do gerenciamento de risco definidos por PMI [6], se mostram de fundamental importância para projetos de implementação de ERP e serão foco deste estudo.

A análise de riscos, que também referenciada por alguns autores como avaliação de riscos, trata-se de analisar os riscos obtidos no processo de identificação de riscos e decidir qual ou quais riscos devem garantir uma resposta no próximo passo [10]. Para ser efetivo, um método de análise de riscos de projeto deve considerar diversos aspectos potenciais, tais como tecnológicos,

mercadológicos, financeiros, operacionais, organizacionais e de negócio [2]. Outra característica importante a ser considerada pelo método de análise de riscos é a sazonalidade, que deve ser considerada indiretamente associando-se a estes aspectos potenciais, ou diretamente quando influencia o evento causador do risco.

Segundo Passari [13], a sazonalidade define-se como flutuações periódicas que apresentam um padrão de longo prazo constante e que, por exemplo, repetem anualmente, semestralmente ou trimestralmente. Passari [13] também cita que ”um aquecimento da economia próximo ao fim do ano” pode ser considerado como um exemplo de sazonalidade. Retomando os aspectos apresentados por Aloini, Dulmin e Mininno [2], é possível inferir que a sazonalidade causada pelo aquecimento da economia exemplificado por Passari [13] poderia impactar o aspecto financeiro e portanto ser considerado para análise de riscos de projetos. Considerando-se também que a sazonalidade pode impactar diretamente os eventos causadores do risco, ACEBES et al. [1] apresenta um outro exemplo em que o período de inverno é estudado como uma incerteza sazonal, que ameaça uma das atividades de um projeto devido ao risco de incidência de temperaturas abaixo de zero grau Celsius. Para Sistemas de Informação (SI), é possível demonstrar um exemplo de sazonalidade através do *freezing*, que conforme apresentado por Neubarth et al. [12], trata-se dos períodos em que somente mudanças emergenciais podem ocorrer no ambiente de tecnologia e que, no caso das instituições financeiras, acontece nos meses de novembro e dezembro devido a necessidade de processamento das informações de fechamento contábil e financeiro pelas empresas. Apesar da influência sazonal relacionada ao *freezing*, Neubarth et al. [12] não apresenta o assunto como sazonalidade.

Com a utilização de métodos de análise de riscos que considerem incertezas sazonais em projetos de implementação de ERP, seria possível mensurar melhor a probabilidade e impacto dos riscos associados a estas incertezas e fornecer informações mais precisas para uma melhor tomada de decisão.

O restante deste artigo está organizado da seguinte forma. Na Seção 2 é apresentado o problema de pesquisa e na Seção 3 a solução proposta para resolver o problema em questão. A Seção 4 descreve o projeto para de avaliação da solução. Na Seção 5, são relatadas as atividades já realizadas. Para concluir, resume-se, na seção 6, as principais contribuições deste artigo.

2. APRESENTAÇÃO DO PROBLEMA

Os projetos de implantação de ERP possuem muitos riscos e o gerenciamento de riscos suporta: melhores decisões sobre planejamento e desenho de processos para prevenir ou evitar riscos; melhor planejamento de contingências para lidar com riscos e seus impactos; e melhor alocação de recursos e orçamento para os riscos [8]. Entretanto, os métodos de avaliação de riscos formais são raramente aplicados para o gerenciamento de riscos de projetos de TI complexos, tais como os projetos de implementação de sistemas de ERP [2]. Segundo Globerson e Zwikael [4], os gerentes de projeto possuem uma lacuna de conhecimento em relação as ferramentas e técnicas formais para o planejamento do gerenciamento de riscos em projetos.

Por outro lado, aspectos que influenciam os projetos e seus riscos, assim como a sazonalidade, deveriam ser considerados por profissionais na análise de riscos de projetos de implementação de ERP. Entretanto, quando o assunto está relacionado a análise de riscos em projetos com influência de incertezas sazonais, não foram encontrados estudos e trabalhos relacionados a área de SI. Além disso, em relação aos trabalhos relacionados a análise de riscos em projetos, a maioria das pesquisas e trabalhos encontrados na literatura não estão relacionadas a SI, mas principalmente as áreas de engenharia e ciências ambientais.

3. PROPOSTA DE SOLUÇÃO

Este trabalho caracteriza-se por ser uma pesquisa qualitativa com o objetivo de identificar na literatura os principais métodos de análise de riscos utilizados para tratar riscos em projetos de implementação de ERP. Esta relação de métodos funcionará como um guia, no qual os profissionais e os estudiosos da área possam ter acesso e, então, definir quais são os métodos mais adequados para o analisar riscos em projetos com influência de sazonalidade, alcançando assim o objetivo geral deste trabalho.

Será utilizado neste estudo a técnica Delphi, que segundo Linstone e Turoff [9], pode ser caracterizada por um processo de comunicação em grupo estruturado, para que seja efetivo ao permitir que o grupo como um todo consiga lidar com um problema complexo. O processo de comunicação é possível através de contribuições de informações e conhecimentos pelos indivíduos, oportunidade de revisão das escolhas pelos indivíduos e pelo grau de anonimato entre as respostas dos indivíduos [9].

A técnica Delphi envolve uma seleção de especialistas baseada em critérios pré-estabelecidos e múltiplas rodadas de questionamento a estes especialistas (através de questionário ou entrevista), aplicadas individualmente de forma a evitar o confronto direto entre eles [3]. Segundo Linstone e Turoff [9], a repetição do questionamento permite um retorno sobre as informações coletadas do grupo nas rodadas anteriores e a oportunidade dos indivíduos de modificarem e refinarem seus julgamentos. Por outro lado, a aplicação do questionário separadamente aos participantes fornece um alto grau de individualidade nas opiniões [9].

De acordo com Skulmoski, Hartman e Krahn [16], o método Delphi é adequado principalmente para estudos onde o objetivo é melhorar o entendimento sobre problemas, oportunidades ou soluções. Além disso, os autores complementam que esse método de pesquisa demonstra-se ser flexível e apropriado para várias necessidades de SI.

Na área de gerenciamento de riscos de projetos de SI, o método Delphi tem sido utilizado principalmente para priorizar os fatores de risco envolvidos nesses projetos [5, 15, 11, 7].

4. PROJETO DE AVALIAÇÃO DA SOLUÇÃO

A identificação do método adequado para analisar riscos em projetos com influência de sazonalidade está relacionada com a execução das quatro fases detalhadas a seguir:

- Na **primeira fase** será identificado, através de uma revisão sistemática da literatura, métodos utilizados para análise de riscos em projetos;
- Na **segunda fase** será elaborado um instrumento com base nos métodos de análise de riscos levantados na literatura e listados em um modelo de referência da pesquisa, para aplicação da técnica Delphi;
- Na **terceira fase** será identificado, com o apoio de especialistas, os métodos mais adequados para se analisar riscos considerando incertezas sazonais em projetos de implementação de ERP, aplicando-se o instrumento elaborado na segunda fase e analisando-se os resultados obtidos; e

- Na **quarta fase** serão ordenados os métodos de análise de riscos conforme a importância atribuída pelos especialistas. Nesta fase serão também elaboradas as considerações finais da pesquisa.

5. ATIVIDADES JÁ REALIZADAS

Já foi realizada a revisão da literatura com objetivo de identificar os conceitos relevantes na área de gerenciamento de riscos, avaliando os principais frameworks utilizados na literatura para gerenciamento de riscos em projetos. Com base na revisão literária, foi iniciada uma revisão sistemática da literatura para identificar métodos utilizados para análise de riscos em projetos, que servirão como um guia para os profissionais e os estudiosos da área definirem quais são os mais adequados para se considerar a influência das incertezas sazonais.

6. CONCLUSÃO

Em geral, o estudo que será realizado pretende abordar os aspectos da sazonalidade em projetos de implementação de ERP, devido aos efeitos causados nos riscos deste tipo de projeto e por serem aspectos pouco explorados pela literatura. O trabalho apresentado por ACEBES et al. [1] com simulações de Monte Carlo, associado ao conceito de *freezing* apresentado por Neubarth et al. [12], mostram a importância de se considerar estes aspectos de sazonalidade ao analisar os riscos em projetos de implementação de ERP, visto que o planejamento ou replanejamento de atividades considerando os riscos influenciados pela sazonalidade é de suma importância para estes sejam evitados ou que os impactos sejam minimizados. Outra questão a ser abordada é a importância do gerenciamento de risco nos projetos de implantação de ERP, visto os métodos de avaliação de riscos formais são raramente aplicados segundo Aloini, Dulmin e Mininno [2].

7. REFERÊNCIAS

[1] F. Acebes, J. Pajares, J. M. Galán, and A. López-Paredes. Exploring the influence of seasonal uncertainty in project risk management. *Procedia - Social and Behavioral Sciences*, 119:329-338, 2014.

[2] D. Aloini, R. Dulmin, and V. Mininno. Risk management in {ERP} project introduction: Review of the literature. 44(6):547-567, 2007.

[3] N. Dalkey and O. Helmer. An experimental application of the delphi method to the use of experts. Management science, 9(3):458-467, 1963.

[4] S. Globerson and O. Zwikael. The impact of the project manager on project management planning processes. Project management journal, 33(3):58-64, 2002.

[5] S.-M. Huang, I.-C. Chang, S.-H. Li, and M.-T. Lin. Assessing risk in erp projects: identify and prioritize the factors. *Industrial management & data systems*, 104(8):681-688, 2004.

[6] P. M. Institute. A Guide to the Project Management Body of Knowledge: PMBOK Guide. PMBOK ® Guide Series. Project Management Institute, 2013.

[7] M. Keil, P. E. Cule, K. Lyytinen, and R. C. Schmidt. A framework for identifying software project risks. *Commun*. ACM, 41(11):76-83, Nov. 1998.

[8] Y. Kop, H. Z. Ulukan, and T. Gürbüz. Evaluating the failure risk level of an enterprise resource planning project using analytic network process in fuzzy environment. *Journal of Multiple-Valued Logic & Soft Computing*, 17(4), 2011.

[9] H. A. Linstone and M. Turo. The delphi method. *Techniques and applications*, 53, 2002.

[10] R. Mulcahy. Risk Management: Tricks of the Trade ® for Project Managers: a Course in a Book *[trademark Symbol]*. RMC Pub., 2003.

[11] R. T. Nakatsu and C. L. Iacovou. A comparative study of important risk factors involved in o shore and domestic outsourcing of software development projects: A two-panel delphi study. *Information & Management*, 46(1):57-68, 2009.

[12] R. Neubarth, E. Guedes, E. de Araújo, and A. Rosini. Governança de ti e gestão de mudanças: Impactos sobre o aumento da disponibilidade nas plataformas de negócio de uma instituição financeira, 2016.

[13] A. F. L. Passari. Exploração de dados atomizados para previsão de vendas no varejo utilizando redes neurais. PhD thesis, *Universidade de São Paulo*, 2003.

[14] P. Rajagopal. An innovation – diffusion view of implementation of enterprise resource planning (erp) systems and development of a research model. *Information and Management*, 40(2):87-114, 2002.

[15] R. Schmidt, K. Lyytinen, and P. C. Mark Keil. Identifying software project risks: An international delphi study. *Journal of management information systems*, 17(4):5-36, 2001.

[16] G. J. Skulmoski, F. T. Hartman, and J. Krahn. The delphi method for graduate research. *Journal of information technology education*, 6:1, 2007.

[17] Standish Group. Big bang boom, Boston, MA: The Standish Group International, Inc. 2014.

[18] W.-H. Tsai, S.-J. Lin, W.-R. Lin, and J.-Y. Liu. The relationship between planning & control risk and erp project success. In *Industrial Engineering an Engineering Management*, 2009. *IEEM* 2009. *IEEE International Conference on*, pages 1835-1839. IEEE, 2009.

Proposta de Modelo de Maturidade para Segurança da Informação baseada na ISO/IEC 27001 e 27002 aderente aos Princípios da Governança Ágil

Alternative Title: Proposal of Information Security Maturity Model based on ISO/IEC 27001 and 27002 according to the Principles of Agile Governance

Gliner Dias Alencar
Centro de Informática (CIn)
Universidade Federal de Pernambuco (UFPE)
Recife, Pernambuco, Brasil.
gda2@cin.ufpe.br

Hermano Perrelli de Moura
Centro de Informática (CIn)
Universidade Federal de Pernambuco (UFPE)
Recife, Pernambuco, Brasil
hermano@cin.ufpe.br

RESUMO

A adoção de um modelo para gestão da segurança da informação, implementação de políticas e adequação a alguma norma de segurança da informação não é algo simples, consequentemente, tem-se dificuldades em sua implantação devido, muitas vezes, a complexidade das normas. O que demonstra a necessidade de pesquisar formas para tentar suprir esta carência. Para isso, este artigo propõem um modelo de maturidade para a segurança da informação como base nos princípios expostos nas normas ISO/IEC 27001 e 27002 e na Governança Ágil. A pesquisa proposta realizará uma revisão sistemática da literatura e questionários para levantamento da situação atual da área de Segurança da Informação nas empresas e dos principais controles necessários. A validação e refinamentos serão obtidos com base em questionários enviados a empresas e a especialistas na área e Grupo Focal.

Palavras-Chave

Segurança da Informação, Modelo de Maturidade, Governança de TIC, Governança Ágil.

ABSTRACT

The adoption of a model for information security management, along with the implementation of its policies and the required adjustments to some of its norms are not simple tasks. Therefore, the implementation of a model for information security management often implies in difficulties due to the complexity of the norms. These difficulties demonstrate the need for a research focused on new ways to overcome such deficiency. To achieve this goal, this paper proposes a maturity model for information security based on the principles exposed on the ISO/IEC 27001 and 27002 Standards and Agile Governance. The proposed research will realize a systematic review of the literature and surveys regarding the current situation of information security in the industry and the main controls currently required. The validation and refinement will be obtained by relying on surveys sent to companies and experts in the area and Focus Group.

SBSI 2017, June 5th–8th, 2017, Lavras, Minas Gerais, Brazil.

Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection.

General Terms

Security.

Keywords

Information Security, Maturity Model, IT Governance, Agile Governance.

1. INTRODUÇÃO

Na sociedade atual, globalizada, competitiva e que necessita de ações e tomadas de decisões rápidas e alinhadas ao negócio, a obtenção e a guarda do conhecimento é de suma importância. Neste contexto, a informação tornou-se um dos mais valiosos ativos das empresas, visto que as informações manuseadas nas corporações podem gerar tanto lucro como grandes prejuízos, assim como o setor que a gerencia tornou-se, nas organizações, estratégico [9].

Mesmo sabendo da importância das informações e da criticidade dos riscos atuais, diversas organizações não contam com planos adequados na área de segurança da informação e alinhamento dos mesmos ao negócio. Em alguns casos, as organizações adotam medidas de Segurança da Informação apenas para atender as forças externas, normalmente oriundas de obrigações legais e regulamentares [4].

1.1 Motivação e Justificativa

O aumento dos incidentes de segurança cresce aceleradamente em todo o mundo. Os ataques atingem diversos tipos de organizações, tanto as governamentais quanto empresas privadas de diversos portes e segmentos. Além disso, vem se tornando cada vez maior a lista de empresas, países e instituições

governamentais que estão em um verdadeiro duelo contra "hackerativistas" [17].

Por conta deste e de outros fatores, existem diversos padrões, frameworks, normas e regulamentos para a implementação de modelos de segurança. Eles fornecem diretrizes ou um conjunto de boas práticas visando a Gestão da Segurança da Informação que, em sua maioria, para incorporar todos os possíveis pontos inerentes à Segurança da Informação, torna-se grande e complexo, fazendo com que, de forma geral, as empresas não os apliquem adequadamente e não gerenciem as características de segurança da informação de forma adequada.

A complexidade e formalismo dos modelos tradicionais mais utilizados atualmente abre oportunidade para rever os processos de implantação de tais padrões, modelos, normas ou frameworks, adequando-os às necessidades específicas de cada organização, visto que mesmo não implantando todos os processos ou controles, a organização consegue obter uma grande mudança organizacional, melhoria em seus processos e maior alinhamento entre a área de TIC e as estratégias organizacionais [16][19].

Almeida Neto et al. [7] também ressaltam esse problema ao apontar a necessidade de se ter um maior controle nas empresas. Porém, é necessário ter agilidade para tratar essas questões no cenário dinâmico atual.

É importante destacar, também, a escassez de estudos que investiguem a presente temática, comparando com outras áreas da computação. Por exemplo, tem-se constantemente visto nos tópicos de interesses dos últimos anos dos principais eventos da área (pode-se citar: SBSI – Simpósio Brasileiro de Sistemas de Informação; CONTECSI – Congresso Internacional de Gestão da Tecnologia e Sistemas de Informação; SBSEG – Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais; e o SBTI – Simpósio Brasileiro de Tecnologia da Informação) chamadas para a área de "Governança de TIC", "Gestão da Segurança da Informação", "Normatização da Segurança da Informação", "Políticas de Segurança da Informação", "Maturidade em Segurança da Informação", ou temas semelhantes, porém, ainda são poucos os trabalhos que abordem tais áreas nos anais.

Neste contexto, acredita-se ser relevante para a área de segurança da informação realizar estudos que busquem produzir modelos para aferir a maturidade da área de segurança da informação. Bem como, dar subsídios para um melhor alinhamento da área à Governança Ágil de TIC e ao negócio buscando meios menos complexos ou burocráticos que os atuais. Com este pensamento, espera-se que uma visão sistêmica da área de segurança da informação com processos menos burocráticos e complexos modifique, de forma positiva, o ambiente corporativo de maneira geral.

2. APRESENTAÇÃO DO PROBLEMA

A ISO (International Organization for Standardization) criou a Família de normas 27000 que versam sobre a segurança da informação. Sendo esse um dos principais mecanismos na área de segurança da informação no que tange, especialmente, aos aspectos táticos e operacionais. Apesar destes modelos serem muito bem estruturados, o formalismo, por algumas vezes excessivo, tem tornado a adoção e melhoria contínua de seus processos uma tarefa complexa [16][19].

A família de normas ISO de segurança da informação apresenta um conjunto de controles. Porém, estudos que demonstrem se sua aplicação realmente afeta a maturidade de segurança da informação da corporação, assim como os alinhando à Governança Ágil de TIC ainda são escassos.

Formas de mensuração da governança de TIC de uma corporação têm sido exploradas [7] como meios de se analisar a situação da instituição, assim como possibilitando a comparação dos níveis de governança entre corporações distintas. Tal aspecto pode ser útil para agregar valor à empresa, como pode ser visto em casos de análises para mercado de ações, vendas, fusões, etc. A área de engenharia de software é outro exemplo, pois utiliza com constância níveis de qualidade e maturidade para diferenciar empresas e produtos.

Diante do contexto citado, esta pesquisa pretende explorar a área respondendo o questionamento: "De que forma é possível mensurar a Segurança da Informação corporativa alinhando aos princípios da Governança Ágil de TIC?".

3. PROPOSTA DE SOLUÇÃO

O objetivo principal deste trabalho é formular um modelo de maturidade para a área de Segurança da Informação alinhado aos princípios da Governança Ágil de TIC com base nas ISO/IEC 27001 e 27002.

No caminho para atingir o objetivo supracitado, espera-se que outros resultados intermediários surjam e possam auxiliar a melhoria da segurança da informação corporativa. O referido objetivo geral pode ser desdobrado em objetivos específicos, que direcionam os possíveis resultados intermediários, onde se destacam os seguintes:

• Compreender como a segurança da informação é tratada, nos diversos aspectos que abrangem essa área, no meio corporativo;

• Prover a junção dos princípios de Governança Ágil de TIC com a área de Segurança da informação, diminuindo a burocracia e formalismos dos modelos atuais;

• Propor métodos de acompanhamento e análise das ações de segurança da informação;

• Criar um guia simplificado para aplicação de uma Política de Segurança da Informação;

• Elencar os fatores críticos de sucesso para melhoria da Segurança da Informação Corporativa;

• Mensurar o nível e maturidade da segurança da informação na corporação;

• Auxiliar os gestores na tomada de decisão através de um modelo integrado de segurança da informação com a estratégia da empresa.

4. PROJETO DE AVALIAÇÃO DA SOLUÇÃO

Para atender aos objetivos propostos, a pesquisa em questão é categorizada como Exploratória e Descritiva, utilizando os procedimentos técnicos de Pesquisa Bibliográfica, Revisão Sistemática da Literatura, Survey e Grupo Focal. Sendo Quanti-qualitativa e concentrada na área de Ciência da Computação: Sistemas de Informação e Segurança da Informação.

A execução da pesquisa será realizada em atividades divididas em duas fases distintas, conforme Figura 1. A primeira etapa consiste em uma fase exploratória e tem por objetivo a construção de uma base teórica consistente para suportar a etapa seguinte. A etapa 2, apresenta uma característica mais descritiva e visa a real construção do modelo proposto.

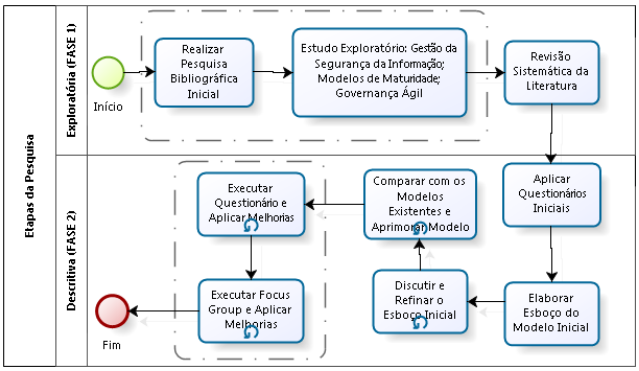


Figura 1. Etapas da Pesquisa

A Revisão Sistemática da Literatura é baseada no método de Kitchenham [11] em bases nacionais e internacionais. Esta etapa ainda encontra-se em andamento com o objetivo de analisar os trabalhos publicados até junho de 2017.

A fase de aplicar questionário inicial consiste no levantamento de características das empresas participantes e da sua visão quanto a importância de cada controle das ISO 27001 e 27002, para isso foram utilizados dois surveys.

O primeiro survey é uma aplicação do documento já utilizado outras vezes [5][6] e que baseou o questionário utilizado por Silva Neto et al. [19]. Sendo composto por 43 questões divididas em seis categorias, sendo elas: Dados da empresa; Dados do respondente; Importância estratégica da informação; Ferramentas de SI na empresa; Recursos humanos e estrutura organizacional; e Segurança da informação corporativa. Ao final foram adicionadas mais cinco questões inerentes a amplitude da pesquisa atual.

O segundo, entregue na mesma etapa, consta do nome da empresa (para correlacionar com o primeiro) e os 114 controles da versão de 2013 da ISO/IEC 27001 e 27002. Para os controles foram utilizados uma escala likert de cinco níveis em uma escala de 1 (nenhuma importância) a 5 (muito importante), sendo a nota 3 categorizada como neutro na escala. Nas questões o respondente marcará a importância de cada controle para o seu ambiente. Mesmo método utilizado por Silva Neto et al. [19], porém os autores colocavam a versão anterior da ISO/IEC 27002 (de 2005) com 133 controles.

A amostra deverá ser composta por empresas de todas as regiões do Brasil, bem como com abrangência de atuação local, regional, nacional e multinacional, sendo contabilizada apenas uma resposta por empresa. Os respondentes deverão ser da área de TIC e, preferencialmente, o responsável pela área de segurança da informação.

Após a análise, os resultados dos controles serão ordenados de acordo com sua média das notas de importância. Posteriormente verificado se existe algum pré-requisito entre os controles e, caso exista, os pré-requisitos serão inseridos antes. Essa fase gerará o Esboço do Modelo Inicial que, posteriormente, será debatido e refinado. Após superada esta etapa, será comparado com os modelos de maturidade existentes e ajustado. Estas etapas poderão ser repetidas.

O modelo refinado e ajustado será enviado aos especialistas na área, procurando obter, ao menos, 5 respostas, e também enviado às empresas respondentes junto com o survey 3. Este terceiro survey questiona sobre a sua adequação, sendo a primeira etapa de validação. Após receber as respostas dos especialistas e empresas, será analisada e aplicada as melhorias propostas. Esta etapa poderá ser repetida.

Tendo o modelo pré-validado e melhorado pela indicação das empresas e especialistas, será realizado grupo focal para a validação final do mesmo. Esta etapa, semelhante a validação do modelo de maturidade de Almeida Neto et al. [8], poderá ser repetida.

Desta forma, acredita-se que objetivo principal seja atingido pela criação, ao final do trabalho, do modelo de maturidade para segurança da informação aderente aos princípios da governança ágil.

Esperar-se, também, que os demais objetivos específicos propostos, não atendidos com o modelo de maturidade formulado, sejam solucionados com as etapas intermediárias para a construção do trabalho e do modelo de maturidade: a análise dos dados coletados e da revisão da literatura. Como consequência dos passos realizados e da solução dos objetivos propostos, acredita-se que o problema de pesquisa seja resolvido a contento.

5. BASES TEÓRICAS E TRABALHOS CORRELATOS

A presente pesquisa baseia-se, principalmente, nos conhecimentos das áreas de Governança de Segurança da Informação [3][15], Governança Ágil [12][13], Modelos de Maturidade [10] [18] e normas ISO/IEC 27001 [1] e 27002 [2].

Na literatura percebe-se um conjunto de trabalhos que tratam de melhorias e maturidade para segurança da informação que podem ser inseridos como correlatos, entre os trabalhos pode-se citar: Rigon et al. [18], Karokola et al. [10] e Mahopo et al. [14]. Porém eles ainda sofrem com os já citados problemas de burocracia e formalismo.

O modelo de maturidade de Almeida Neto et al. [7][8] tenta solucionar os problemas da burocracia e formalismo, baseando-se nos princípios da Governança Ágil, porém não tem foco em segurança da informação. Já Silva Neto et al. [19], propõe uma simplificação da segurança, mas apresenta apenas uma visão inicial de uma política de segurança da informação simplificada, não formulando um modelo de maturidade. Assim, acredita-se que os trabalhos encontrados não atendem, por completo, todos os objetivos específicos propostos nesta presente pesquisa e problemas expostos.

6. ATIVIDADES JÁ REALIZADAS

Dentro das etapas propostas no método, já foram realizadas: a Pesquisa Bibliográfica Inicial; o Estudo Exploratório das áreas envolvidas. A Revisão Sistemática da Literatura já foi iniciada e

continua em andamento objetivando inserir os trabalhos publicados até junho deste ano (2017.1).

Os questionários iniciais já foram aplicados, sendo enviado para 341 empresas (todos os respondentes serão denominados como "empresa", indiferente da classificação, porte ou abrangência), sendo respondidos por 229 (67,2%). Destas, apenas 157 respostas foram consideradas em conformidade com os objetivos da pesquisa (46% do total enviado e 68,6% das respostas obtidas), ou seja, 157 empresas distintas e que responderam todas as perguntas.

Com relação à Economia das Empresas alcançadas, tem-se: 83% Privada, 3% Mista e 14% Pública. Com relação à Atuação: 28% Local, 24% Regional, 33% Nacional e 15% Multinacional. Sendo 21% com menos de 50 funcionários, 27% entre 50 e 100, 26% entre 100 e 200, 14% entre 200 e 500 e, por fim, 12% com mais de 500 funcionários. Vale ressaltar que se alcançou empresas de todas as regiões do Brasil.

O Modelo Inicial já foi elaborado. Este modelo já foi discutido e refinado, bem como comparado inicialmente com outros modelos existentes e um questionário de validação com empresas e com especialistas já foi aplicado. Conforme método proposto (Figura 1), essas fases de debate, comparação e aplicação de questionários para refinamento do modelo podem se repetir. Estando o trabalho nesta fase.

7. CONSIDERAÇÕES FINAIS

Espera-se que o modelo proposto promova apoio significativo à adoção e melhorias contínuas à Segurança da Informação mensurando a maturidade da empresa; apontando as áreas com maior desenvolvimento em segurança da informação e as áreas que precisam de maior investimento; ter, de forma palpável, o impacto na segurança da informação de alterações (sejam elas pessoais, procedimentais ou tecnológicas) ocorridas na corporação; possibilidade de comparação do "nível de maturidade de segurança" entre setores ou empresas; diminuição da burocracia e formalismo na área gerando maior agilidade na gestão da segurança da informação corporativa.

Assim, acredita-se que o presente trabalho resultará em contribuições significativas para a área de Segurança da Informação, o que deverá resultar em melhorias para todas as áreas de TIC. Assim como, as contribuições que se pretende galgar serão aplicáveis tanto no contexto acadêmico quanto no corporativo dando subsídios para que as empresas melhorem seu "nível de segurança da informação" e consigam atuar de forma mais competitiva no atual mercado globalizado.

8. REFERÊNCIAS

[1] ABNT NBR ISO/IEC 27001. 2013. NBR ISO/IEC 27001 - Sistema de Gestão de Segurança da Informação – Requisitos.

[2] ABNT NBR ISO/IEC 27002. 2013. NBR ISO/IEC 27002 - Tecnologia da informação - técnicas de segurança - código de prática para a gestão da segurança da informação.

[3] ABNT NBR ISO/IEC 27014. 2013. NBR ISO/IEC 27014 - Tecnologia da Informação – Técnicas de Segurança – Governança de segurança da informação.

[4] Albuquerque Junior, A. E., and Santos, E. M. 2014. Adoção de medidas de Segurança da Informação: um modelo de análise para institutos de pesquisa públicos. *Revista Brasileira de Administração Científica*, 5, 2 (2014), 46-59. DOI= http://dx.doi.org/10.6008/SPC2179-684X.2014.002.0004

[5] Alencar, G. D., Queiroz, A. A. L., and de Queiroz, R. J. G. B. 2013. Insiders: Análise e Possibilidades de Mitigação de Ameaças Internas. *Revista Eletrônica de Sistemas de Informação*, *12,* 3, artigo 6 (set-dez 2013), 38 páginas.

DOI= http://dx.doi.org/10.5329/RESI.2013.1203006

[6] Alencar, G. D., Queiroz, A. A. L, and de Queiroz, R. J. G. B. 2013. Insiders: Um Fator Ativo na Segurança da Informação. SBSI - Simpósio Brasileiro de Sistemas de Informação. In *Anais do IX Simpósio Brasileiro de Sistemas de Informação* (João Pessoa – PB). SBSI'13, SBC. 254-259.

[7] Almeida Neto, H. R., de Magalhães, E. M. C., de Moura, H. P., de Almeida Teixeira Filho, J. G., Cappelli, C., and Martins, L. M. F. 2015. Avaliação de um Modelo de Maturidade para Governança Ágil em Tecnologia da Informação e Comunicação. *iSys-Revista Brasileira de Sistemas de Informação*, 8, 4 (2015), 44-79.

[8] Almeida Neto, H. R., Magalhães, E. M. C., Moura, H. P., Teixeira Filho, J. G. A., Capelli, C., and Martins, L. M. F. 2015. Avaliação de um Modelo de Maturidade para Governança Ágil em TIC usando Focus Group. In *Anais do XI Simpósio Brasileiro de Sistemas de Informação* (Goiânia – GO). SBSI'15, SBC. 15-22.

[9] Castells, M. 2007. *Era da Informação: A Sociedade em Rede,* Volume 1, 10ª Edição. São Paulo: Editora Paz e Terra.

[10] Karokola, G., Kowalski, S., and Yngström, L. 2011. Towards An Information Security Maturity Model for Secure e-Government Services: A Stakeholders View. In *Proceedings of the Fifth International Symposium on Human Aspects of Information Security and Assurance*, (Londres – Inglaterra), HAISA'11. 58-73.

[11] Kitchenham, B. 2004. *Procedures for performing systematic reviews*. Technical Report. Keele University.

[12] Luna, A. J. D. O., Kruchten, P., Pedrosa, M. L. D. E., Neto, H. R., and De Moura, H. P. 2014. State of the art of agile governance: a systematic review. *International Journal of Computer Science and Information Technology*, 6, 5 (2014), 121–141. DOI = http://dx.doi.org/10.5121/ijcsit.2014.6510

[13] Luna, A. J. H. O.; Kruchten, P.; Riccio, E. L., and De Moura, H. P. 2016. Foundations For An Agile Governance Manifesto: A Bridge For Business Agility. In *Proceedings of the 13th International Conference on Information Systems and Technology Management* (São Paulo – SP), CONTECSI'16, USP. 4391-4404.

[14] Mahopo, B., Abdullah, H., and Mujinga, M. 2015. A formal qualitative risk management approach for IT security. In *Information Security for South Africa* (Joanesburgo - África do Sul), ISSA'15*,* IEEE. 1-8.

[15] Manoel, S. S. 2014. *Governança de Segurança da Informação: Como criar oportunidades para o seu negócio*. Rio de Janeiro: Editora Brasport.

[16] Prado, E. P. V., Mancini, M., Barata, A. M., and Sun, V. 2016. Governança de TI em Organizações do Setor de Saúde: um Estudo de Caso de Aplicação do COBIT. In *Anais do XII Simpósio Brasileiro de Sistemas de Informação* (Florianópolis – SC), SBSI'16, SBC. 1-8.

[17] Pwc. 2016. *PricewaterhouseCoopers. Pesquisa global de segurança da informação 2016*. http://www.pwc.com.br/pt/publicacoes/servicos/consultoria-negocios/2016/pwc-pesquisa-global-seguranca-informacao-16.html

[18] Rigon, E. A, Westphall, C. M, Santos, D. R., and Westphall, C. B. 2014. A cyclical evaluation model of information security maturity. *Information Management & Computer Security*, 22, 3 (2014), 265-278. DOI= http://dx.doi.org/10.1108/IMCS-04-2013-0025

[19] Silva Neto, G. M.; Alencar, G. D., and Queiroz, A. A. L. 2015. Proposta de Modelo de Segurança Simplificado para Pequenas e Médias Empresas. In *Anais do XI Simpósio Brasileiro de Sistemas de Informação* (Goiânia – GO), SBSI'15, SBC. 299-306.

Um Processo para Gerenciamento de Requisitos de Sistema de Sistemas
Alternative Title: A Management Process for System of Systems Requirements

Renata Martinuzzi de Lima
Programa de Pós-graduação
em Ciência da Computação
Universidade Federal de Santa Maria (UFSM)
Santa Maria
Rio Grande do Sul - Brasil
rlima@inf.ufsm.br

Lisandra Manzoni Fontoura
Programa de Pós-graduação
em Ciência da Computação
Universidade Federal de Santa Maria (UFSM)
Santa Maria
Rio Grande do Sul - Brasil
lisandra@inf.ufsm.br

RESUMO
Sistema de Sistemas (SoS) estão se tornando cada vez mais comuns em nossa sociedade global, por isso há um crescente interesse em fechar as lacunas ainda existentes no âmbito da Engenharia de Sistema de Sistemas (ESoS), campo que ainda encontra-se em constante crescimento devido a sua complexidade. Grande parte dessas lacunas relacionam-se a Engenharia e Gerenciamento dos requisitos de SoS. A ausência de métodos e técnicas bem definidas e padronizadas para a Engenharia de Requisitos de SoS dificulta o processo de desenvolvimento e a evolução do SoS como um todo. Portanto, este trabalho visa propor um processo de gerenciamento de requisitos de SoS capaz de organizar o trabalho colaborativo entre os diversos *stakeholders*, além de antecipar e gerenciar mudanças nos requisitos, colaborando assim para um desenvolvimento evolucionário mais previsível e adaptável.

Palavras-Chave
Sistema de Sistemas; Engenharia de Requisitos; Gerenciamento de Requisitos.

ABSTRACT
System of Systems (SoS) are becoming more common in our global society. Therefore, there is an increasing interest in closing the gaps still existing in the System of Systems Engineering (SoSE) context, which is a field in constant growth due to its complexity. Most of these gaps are related to the SoS requirements engineering or management. The lack of well-defined and standardized methods and techniques for Requirements Engineering for SoS (RESoS) makes difficult the development process and the evolution of the overall SoS. Hence, in this paper we propose a requirements management process for SoS capable of organizing the collaborative work among stakeholders, besides anticipating and managing changes in requirements, thus contributing to a more predictable and adaptable evolutionary development of the SoS.

SBSI 2017, June 5th–8th, 2017, Lavras, Minas Gerais, Brazil.

Keywords
System of Systems; Requirements Engineering; Requirements Management.

Categories and Subject Descriptors
D.2.1 [**Requirements/Specifications**]*:* Elicitation methods (*e.g., rapid prototyping, interviews, JAD*) and Methodologies.

General Terms
Management, Documentation; Standardization.

1. INTRODUÇÃO
Em diferentes domínios de aplicação, é possível encontrar a atuação de sistemas distintos, geralmente já existentes, operando em conjunto para satisfazer um determinado objetivo. Estes sistemas constituintes, juntos estabelecem o que é denominado de "Sistema de Sistemas" ou pelo termo em inglês "*System of Systems*" (SoS), termo que vem sendo usado para descrever estes sistemas complexos, compostos de sistemas constituintes independentes, os quais trabalhando em conjunto conseguem atingir um objetivo específico através da sua sinergia [7].

A Engenharia de Sistema de Sistemas (ESoS) é um assunto de crescente interesse dentro da comunidade de Engenharia de Sistemas (ES). No entanto, as diferenças estruturais e a grande diversidade de *stakeholders* em um projeto de desenvolvimento de um SoS levam a complexidades e desafios ainda pouco explorados na Engenharia de Sistemas tradicional [9].

Os sete Elementos Centrais da ESoS, definidos por [3], descrevem como processos da ES são aplicados a SoS. Três destes elementos centrais relacionam-se com a Engenharia de Requisitos

de SoS (ERSoS), são eles: *"Translating Capability Objectives", "Developing and Evolving an SoS Architecture", e "Adressing Requirements and Solution Options"*.

Por meio da descrição desses processos é possível se ter uma ideia de como as atividades devem ser executadas, porém não há um padrão definido, nem um método consistente para gerenciamento, elicitação, análise, especificação, verificação ou validação dos requisitos de um SoS. Ou seja, o paradigma da ERSoS ainda não atingiu um nível de maturidade ou sofisticação experienciado pela ER tradicional. Portanto, as abordagens bem sucedidas no âmbito da ERSoS ainda são escassas e muitos problemas e desafios de pesquisa podem ser encontrados [4].

Uma questão recorrente citada em diversos estudos da área na última década, trata-se da necessidade de um gerenciamento de requisitos que facilite o trabalho colaborativo entre o grande número de *stakeholders* [7] e que apoie a evolução do SoS de forma adaptável e flexível [1].

Um processo de gerenciamento de requisitos de SoS bem definido seria capaz clarificar as relações de gerenciamento entre o SoS e os sistemas constituintes facilitando a colaboração entre os diversos envolvidos que participam do ciclo de desenvolvimento de um SoS. Assim, opções de gerenciamento mais consistentes facilitariam a ERSoS, garantindo um processo de desenvolvimento mais previsível e adaptável.

Este artigo apresenta uma proposta de dissertação de mestrado, que tem como objetivo propor um protótipo de processo de gerenciamento de requisitos para a Engenharia de Requisitos de Sistema de Sistemas e está organizado da seguinte forma: A seção dois apresenta o problema de pesquisa, a seção três a proposta de solução, a seção quatro o projeto de avaliação da solução, na seção cinco estão descritas as atividades já realizadas e por fim a conclusão e as referências bibliográficas.

2. APRESENTAÇÃO DO PROBLEMA

DAHMANN, 2016 [1] cita o tópico de capacidades e requisitos como um "ponto fraco" da Engenharia de SoS, pois tradicionalmente, ciclos de desenvolvimento de sistemas começam com um conjunto claro e completo de requisitos e fornecem uma abordagem disciplinada para o desenvolvimento de um sistema que atenda a estes requisitos.

Entretanto, um SoS compreende múltiplos sistemas independentes com seus próprios requisitos que podem ou não alinhar-se aos objetivos de capacidades do SoS [8]. Nesse caso, DAHMANN e BALDWIN [1] aponta a necessidade de clarificar relações de gerenciamento entre o SoS e os sistemas constituintes, garantindo as propriedades-chave do SoS, ou seja, a independência operacional e gerencial de seus sistemas constituintes [7].

Além disso, NIELSEN et al. [7] afirma que devido ao grande número de *stakeholders* envolvidos em um projeto de SoS, incluindo proprietários e operadores dos sistemas constituintes, há uma clara necessidade de se empregar métodos e ferramentas que suportem um trabalho colaborativo para elicitação de requisitos entre estes envolvidos.

Outro ponto importante a ser considerado é que o desenvolvimento de um SoS é um processo longo, contínuo, em constante evolução e susceptível a diversas mudanças nos requisitos. Como declara [6], em um SoS existem muitas oportunidades de mudanças, por isso é possível afirmar que "mudança" é o estado normal de um SoS. Nesse sentido, há também a falta de mecanismos que possam antecipar as mudanças nos requisitos e avaliar seus impactos de forma conjunta entre gerentes e engenheiros dos sistemas constituintes [1].

DAHMANN, 2016 [2] também declara que a falta de uma padronização na Engenharia de Sistema de Sistemas vem se tornando cada vez mais evidente, e assim há a oportunidade de definição de padrões ou adaptação de normas já existentes que possam ser aplicadas no âmbito da ESoS.

3. PROPOSTA DE SOLUÇÃO

Este trabalho propõe desenvolver uma abordagem para gerenciamento de requisitos de Sistema de Sistemas por meio de um processo que seja capaz de organizar o trabalho colaborativo na interação entre os *stakeholders* e que forneça mecanismos flexíveis para a antecipação e adaptação de mudanças nos requisitos durante o processo evolutivo de um SoS.

O processo será definido de modo que consiga integrar técnicas de gerenciamento de trabalho entre engenheiros e gerentes do SoS e dos sistemas constituintes, além de técnicas de antecipação de mudanças nos requisitos, assim clarificando os papéis na tomada de decisões e facilitando a evolução do SoS de forma mais previsível.

O protótipo de processo será modelado usando o metamodelo SPEM 2.0 (*Software & Systems Process Engineering Meta-Model Specification*), notação padrão da OMG (*Object Management Group*) para modelagem de processos. Além disso, uma ferramenta será desenvolvida para apoiar a execução do processo.

4. PROJETO DE AVALIAÇÃO DA SOLUÇÃO

A avaliação e validação da proposta será realizada em duas etapas:

1. A qualidade do processo será avaliada em contraste com as normas e recomendações existentes no âmbito da Engenharia de Sistema de Sistemas e Engenharia de Requisitos: ISO | IEC | IEEE 15288:2015; ISO | IEC | IEEE 29148:2011, além de outras recomendações que venham a ser identificadas durante o estudo teórico. Espera-se que o processo esteja alinhado com as condições sugeridas pelas normas e recomendações identificadas.
2. Para avaliar a viabilidade do processo é planejado um estudo de caso real. O processo será aplicado a um projeto por meio de uma ferramenta de apoio e os dados qualitativos sobre a viabilidade da ferramenta e da proposta de processo serão coletados por meio de questionários e entrevistas com as pessoas envolvidas na definição de requisitos do projeto.

5. ATIVIDADES REALIZADAS

Inicialmente foi realizado um estudo sobre Sistema de Sistemas com o intuito de obter uma visão geral sobre o tema e elencar tópicos de interesse. Após a escolha do tópico

"Engenharia de Requisitos", uma Revisão Sistemática de Literatura (RSL), usando a abordagem *Snowballing*, foi desenvolvida com o objetivo reunir estudos atuais sobre a ERSoS e principalmente encontrar desafios de pesquisa a serem explorados neste tópico.

A Revisão Sistemática de Literatura identificou que entre os desafios de pesquisa mais citados está a necessidade de gerenciamento de requisitos e de suas mudanças ao longo do ciclo de vida do SoS. Além de desafios técnicos relacionados as atividades de validação e verificação de requisitos que considerem as características intrínsecas de um SoS [5].

Definido o problema de pesquisa, com base nos desafios apontados na RSL, um projeto de pesquisa foi desenvolvido a fim de identificar uma proposta de solução.

Atualmente, o trabalho encontra-se em fase de estudo teórico e levantamento bibliográfico focado principalmente nas soluções já existentes e nas normas e recomendações existentes no âmbito da Engenharia de SoS. Após este estudo, dar-se-á início a definição e modelagem do processo de gerenciamento de requisitos para SoS, o desenvolvimento da ferramenta para apoiar o processo e consequentemente a avaliação e validação da viabilidade do mesmo em um estudo de caso.

6. CONCLUSÃO

Este trabalho apresentou a proposta de desenvolvimento de um processo para gerenciamento de requisitos de Sistema de Sistemas. A ideia principal é organizar a engenharia de requisitos de modo que o trabalho entre os gerentes e engenheiros do SoS e dos sistemas constituintes seja colaborativo, ou seja, que eles tenham apoio na definição de requisitos e na incorporação das mudanças que podem surgir nos requisitos de capacidades do SoS ao longo do seu desenvolvimento e evolução.

O trabalho encontra-se em fase de levantamento bibliográfico e estudo teórico das soluções já propostas, e das recomendações para a Engenharia de Requisitos de SoS. Após o desenvolvimento e modelagem do processo, este será avaliado de acordo com as normas existentes no âmbito da Engenharia de Sistema de Sistemas e sua efetividade será validada em um estudo de caso real.

7. AGRADECIMENTOS

Os autores agradecem a Comissão de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio na realização dessa pesquisa.

8. REFERÊNCIAS

[1] DAHMANN, J. S. and Baldwin, K. J. (2008). Understanding the current state of us defense systems of systems and the implications for systems engineering. In Systems Conference, 2008 2nd Annual IEEE, pages 1–7. IEEE.

[2] DAHMANN, Judith; ROEDLER, Garry. Moving towards standardization for system of systems engineering. In: System of Systems Engineering Conference (SoSE), 2016 11th. IEEE, 2016. p. 1-6.

[3] DoD-USA (2008). Systems Engineering Guide for Systems of Systems. Washington, DC.

[4] KEATING, Charles B.; PADILLA, Jose J.; ADAMS, Kevin. System of systems engineering requirements: challenges and guidelines. Engineering Management Journal, v. 20, n. 4, p. 24-31, 2008.

[5] LIMA, Renata M.; VARGAS, Daniel. FONTOURA, Lisandra M.. System of System Requirements: A Systematic Literature Review using Snowballing. In: SEKE: Software Engineering and Knowledge Engineering. Knowledge Systems. Wyndham Pittsburgh University Center, Pittsburgh, USA, 2017. No prelo.

[6] LEWIS, Grace A. et al. Requirements engineering for systems of systems. In: Systems Conference, 2009 3rd Annual IEEE. IEEE, 2009. p. 247-252.

[7] NIELSEN, Claus Ballegaard et al. Systems of systems engineering: basic concepts, model-based techniques, and research directions. ACM Computing Surveys (CSUR), v. 48, n. 2, p. 18, 2015.

[8] WALDEN, David D.; ROEDLER, Garry J.; FORSBERG, Kevin. INCOSE Systems Engineering Handbook Version 4: Updating the Reference for Practitioners. In: INCOSE International Symposium. 2015. p. 678-686.

[9] SEBoK. The Guide to the Systems Engineering Body of Knowledge. V. 1.6. Hoboken, NJ - USA. 2017. Disponivel em: < http://sebokwiki.org/wiki/Guide_to_the_Systems_Engineering_Body_of_Knowledge_(SEBoK) >. Acesso em Março de 2017.

Validação da técnica Business Process Point Analysis (BPPA)

Validation of the Business Process Point Analysis (BPPA) technique

Natália Oliveira
Universidade de São Paulo
Rua Arlindo Béttio, 1000
São Paulo, Brasil
n.oliveira@usp.br

Marcelo Fantinato (orientador)
Universidade de São Paulo
Rua Arlindo Béttio, 1000
São Paulo, Brasil
m.fatinato@usp.br

RESUMO

Em projetos de tecnologia da informação, a medição do tamanho funcional de software representa uma tarefa árdua enfrentada pelos gestores de projetos, que precisam garantir a estimativa correta do tamanho de um projeto a fim de evitar prejuízos à organização. Para esse fim, técnicas sistemáticas de medição para calcular o tamanho funcional de software têm sido usadas na engenharia de software. Considerando que projetos relacionados à gestão de processos de negócio apresentam problemas similares, a técnica *Business Process Points Analysis* (BPPA) foi proposta para possibilitar a estimativa do tamanho de projetos de automação de processos de negócio. BPPA é baseada na técnica *Function Point Analysis* (FPA) da engenharia de software. O projeto de pesquisa de mestrado descrito neste artigo tem o objetivo de validar consistentemente a técnica BPPA, visando avaliar de forma apropriada seus benefícios e suas limitações. Espera-se poder validar se a técnica BPPA apresenta resultados corretos, confiáveis e úteis a gerentes de projeto, considerando uma análise custo/benefício.

Palavras-Chave

Gestão de processos de negócio, métrica, validação.

ABSTRACT

In information technology projects, measuring the functional size of software is an arduous task faced by project managers, who need to ensure the correct estimation of a project size in order to avoid losses to the organization. To this end, systematic measurement techniques for calculating software functional size have been used in software engineering. Considering that projects related to business process management have similar problems, the BPPA technique was proposed to allow the estimation of the size of business process automation projects. BPPA is based on the software engineering's Functional Point Analysis (FPA) technique. The masters research project described in this paper proposes to consistently validate the BPPA technique, aiming to properly evaluate its benefits and limitations. We expect to be able to validate if the BPPA technique presents correct, reliable and useful results to project managers, considering a cost/benefit analysis.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•General and reference → Empirical studies; Measurement; •Applied computing → Business process management;

Keywords

Business process management, metrics, validation.

1. INTRODUÇÃO

A técnica *Business Process Point Analysis* (BPPA), proposta por Baklizky et al. [3] visa estender a técnica *Function Point Analysis* (FPA), da engenharia de software, para ser aplicada em Gestão de Processos de Negócio (BPM – *Business Process Management*). BPPA foi definida sob a supervisão do mesmo orientador deste projeto. Embora uma avaliação de BPPA tenha sido realizada, não foi possível validá-la consistentemente. Assim, não foram validados seus reais benefícios para projetos de BPM, que é seu principal objetivo. Este projeto de pesquisa de mestrado é motivado pela necessidade de realizar tal validação a fim de apresentar à comunidade interessada, informações mais precisas sobre a técnica BPPA.

BPM envolve conceitos, métodos e técnicas para a gestão completa de processos de negócio, incluindo modelagem, implementação e execução. BPM baseia-se na representação explícita de processos, atividades e restrições de execução [9]. É uma abordagem essencial para organizações, pois permite transformar o diálogo entre áreas de negócio e Tecnologia da Informação (TI) em uma interpelação interativa e iterativa [2]. Assim, BPM favorece o alinhamento estratégico entre negócio e TI, gerenciando e aprimorando soluções tecnológicas oferecidas pela TI, a partir de processos que incorporam valor significativo às organizações. BPM pode, portanto, oferecer vantagem competitiva às organizações e aliar a melhoria contínua com a visão estratégica da organização [4].

No contexto similar de engenharia de software, FPA possibilita medir a complexidade funcional de um software com

base em seus requisitos funcionais. FPA é baseada na métrica Pontos de Função (FP – *Function Points*), cujo valor é independente de métodos de desenvolvimento, tecnologias, linguagens e plataforma de hardware a serem usados para projetar e implementar o software [8]. Com base no número de FPs para um sistema de software, calculado por meio de FPA, um gerente de projetos pode derivar diferentes estimativas a serem usadas no projeto de desenvolvimento desse sistema, incluindo custo, tempo, recursos etc. [6].

BPPA foi proposta para que gerentes de projetos da área de BPM possam estimar sistematicamente o tamanho funcional de um processo de negócio a ser usado em um projeto de automação de processo de negócio. Trata-se da primeira técnica proposta para esse fim conforme análise realizada a literatura relacionada [3]. Buscando a definição sólida de uma técnica que oferecesse esse resultado, BPPA foi definida com base em FPA - uma técnica já amplamente usada em engenharia de software para fim similar, porém tendo especificamente como objeto o tamanho funcional de software. Enquanto FPA é baseada na métrica FP, BPPA é baseada na métrica conceitualmente similar BPP (*Business Process Points*). Como resultado da aplicação da técnica BPPA, obtém-se o tamanho funcional de um processo de negócio expresso pelo número de BPPs. O cálculo de BPPs é feito com base em um modelo de processo de negócio expresso via BPMN (*Business Process Model and Notation*).

Embora BPPA tenha recebido uma boa avaliação inicial por um especialista em medição de software e FPA [3], não foi possível validar se BPPA resulta em valores corretos e confiáveis para apoiar gerentes de projeto similarmente ao realizado por FPA. Para que as comunidades científica e industrial observem os reais benefícios de BPPA, é necessário que os resultados de uma validação consistente sejam apresentados, incluindo a apresentação de suas limitações.

Junto a proposta inicial de BPPA, já foi realizada uma avaliação, porém, apenas com o propósito de confirmar a conformidade entre BPPA e FPA, uma vez que BPPA foi baseada conceitualmente em FPA [3]. Assumiu-se que: se FPA oferece benefícios para engenharia de software e se BPPA é equivalente a FPA, então BPPA deveria oferecer benefícios similares a BPM. Para isso, foi realizado um experimento simples em que os participantes aplicaram ambas BPPA e FPA em um cenário simples, para permitir que fossem posteriormente coletadas suas opniões a respeito de cada técnica, principalmente sobre suas semelhanças e diferenças, mesmo considerando que elas são projetadas para contextos distintos. Esse experimento apresentou evidência que as técnicas são similares. Porém, os participantes eram apenas estudantes de graduação de uma disciplina em qualidade de software, que trata de métricas de software.

Durante a análise de BPPA, realizada por avaliadores *ad hoc* (especialistas em BPM ou FPA), várias limitações, ameaças e sugestões foram apresentadas em relação à avaliação inicial de BPPA. Alguns dos pontos mencionados são: A necessidade de tratar fatores como confiabilidade e validade; A obrigatoriedade de comparação entre viabilidade e facilidade de uso da técnica; A inevitabilidade de identificar as limitações (ameaças a validade) da validade da técnica; A indispensabilidade de realizar a validação com recursos como profissionais da área, ou ao menos alunos de pós-graduação; E por fim, ser capaz de identificar a validade a técnica.

Baseado nesse contexto histórico, este projeto de pesquisa de mestrado visa validar consistentemente BPPA, visando avaliar de forma apropriada seus benefícios e suas limitações. Espera-se poder validar BPPA para que se obter maior interesse da comunidade potencialmente interessada.

Como não existem outras propostas na literatura para o mesmo objetivo, não é possível comparar os resultados obtidos com BBPA com outras técnicas similares. Além disso, há poucos trabalhos relacionados que apresentam validações das técnicas propostas, mesmo que para outros tipos de métricas para BPM. Espera-se a identificação desse tipo de trabalho para serem usados como base para a validação pretendida neste projeto de pesquisa.

Em um trabalho de validação similar ao almejado, Abrahão et al. [1] avaliaram um método proposto para a medição de tamanho funcional de esquemas conceituais orientados a objetos. A princípio os autores realizaram uma validação teórica de métricas de software, a fim de demonstrar que o método de medição proposto realmente mede o que é destinado medir, este processo inclui: Definição de unidade de medida; Definição das regras do método de medição; Qualidade do metamodelo; e Solidez das operações matemáticas aplicadas nas regras de medição. Em seguida, os autores realizaram uma auditoria dos resultados da medição, chamada de validação a posteriori, pois a validação a priori, visa rever os procedimentos do processo de coleta de dados.

Os autores aplicaram o Modelo de Avaliação de Métodos (MEM), um modelo teórico para avaliar métodos de projeto de sistemas de informação, que incorpora aspectos duais, permitindo avaliar se o método atinge seus objetivos eficientemente, e se o método será utilizado na prática. Por fim, realizaram a validação de precisão de predição dos modelos descritivos e preditivos. Apesar dos resultados obtidos, os autores apresentam como ameaça à validade da avaliação a não garantia de que o método proposto será sempre aplicado corretamente. Por isso, sugerem um processo de auditoria do resultado da medição durante o processo de medição.

2. PROPOSTA DE SOLUÇÃO

Para validar consistentemente a técnica BPPA, propõe-se realizar um experimento para avaliar de forma apropriada seus benefícios e limitações. É esperado que os resultados apresentem indícios de que a aplicação de BPPA leva a obtenção de resultados corretos e confiáveis. O objetivo do experimento é que os participantes apliquem BPPA sobre processos de negócio para o calcular os BPPs. Com base nos resultados dos cálculos, uma série de análises de corretude e confiabilidade será realizada. A seguir, são apresentadas algumas informações relacionadas ao experimento previsto:

Pretende-se contar com participantes com experiência prévia em BPM. Uma possibilidade são os alunos da disciplina "Gestão automatizada de processos de negócio", a ser ministrada no segundo semestre de 2017, no Programa de Pós-graduação em Sistemas de Informação. São comuns alunos regulares e especiais com experiência profissional, e para a realização desta proposta, espera-se a participação de 30 alunos, onde três modelos de processos de negócio serão utilizados, com diferentes níveis de complexidade. Cada participante deverá aplicar BPPA para cada um dos três modelos, calculando assim três números de BPPs. Para permitir melhores resultados desta aplicação, um treinamento prévio em BPPA será oferecido aos participantes, afim de permitir que conheçam a técnica a ser utilizada.

Com base nos números de BPPs calculados pelos participantes, análises de corretude e de confiabilidade serão

realizadas. Os valores considerados corretos serão calculados previamente pelos pesquisadores proponentes, por serem os únicos especialistas em BPPA. Os proponentes realizarão esse cálculo de forma conjunta, visando a melhor corretude no uso da técnica, a servir como oráculo para o experimento. Técnicas estatísticas descritivas e inferenciais devem ser aplicadas para aumentar a confiabilidade dos resultados obtidos.

Para análise da **corretude**[1], será verificada a proximidade dos números de BPPs calculados pelos participantes em relação ao valor correto. Serão verificados os desvios padrão do valor calculado por cada participante em relação ao valor correto e, ao final, a média de todos os desvios padrão calculados. É esperado uma baixa média de desvios padrão dos participantes em relação ao valor correto.

Para análise da **confiabilidade**, será verificada a proximidade dos valores de BPPs calculados pelos participantes. Serão verificados os desvios padrão do valor calculado por cada participante em relação a média dos valores de todos participantes e, por fim, a média de todos os desvios padrão. É esperado um baixo desvios padrão entre os participantes.

Os cálculos sobre corretude e sobre confiabilidade devem ser realizados três vezes, uma vez para cada modelo de processo considerado, como forma de diminuir as ameaças às validades internas e de construção do experimento.

Como referência para avaliar a corretude e a confiabilidade em termos de desvios padrão, serão usados valores de referência publicados na literatura sobre experimentos similares envolvendo FPA. Valores de referência de FPA precisarão ser usados devido à falta de experiências prévias em BPM.

3. AVALIAÇÃO DA SOLUÇÃO

Será conduzida uma avaliação sistemática das possíveis ameaças à validade do experimento a ser realizado. Essa avaliação considerará diferentes tipos de validade, incluindo, por exemplo [5, 7]: externa/transmissibilidade, interna, construção, credibilidade, confiabilidade e confirmalidade.

Uma possível fonte de ameaça à validade do experimento já observada é em relação ao uso de alunos de pós-graduação em vez de profissionais experientes na indústria. Considerando o objetivo da técnica BPPA em avaliação, seria mais confiável o uso de profissionais com larga experiência em assuntos tais como: BPM, métricas de software e gestão de projetos. Porém, considerando a complexidade do experimento, cuja execução deverá levar várias horas, talvez um dia inteiro de trabalho, torna-se inviável a disponibilidade de tal tipo de profissional. Por isso a escolha por alunos de pós-graduação ligados ao tema BPM é a possível escolha a ser feita. Apesar disso, o experimento está sendo projetado de forma a diminuir os viéses potencialmente existentes nesse contexto. Essa e outras possíveis ameaças à validade do experimento já estão sendo ou ainda serão tratadas.

4. ATIVIDADES JÁ REALIZADAS

Em termos de requisitos formais do curso de mestrado, quase todos os créditos já foram cumpridos nos primeiros dois semestres do curso. Estando agora no terceiro semestre, onde a última disciplina está sendo cursada para a obtenção dos créditos finais necessários para a dissertação.

Sobre o projeto de pesquisa propriamente dito, primeiramente, houve um estudo das teorias básicas relacionadas às áreas de BPM e FPA, incluindo a técnica BPPA, proposta previamente. Na sequência, foi realizada uma análise crítica das sugestões, problemas e críticas apresentadas por revisores *ad hoc*, considerando a proposta original de BPPA e sua avaliação preliminar. Com base nessas duas atividades iniciais, o objetivo deste projeto de pesquisa pode ser definido.

Baseado nos objetivos definidos para este projeto de pesquisa, uma análise bibliográfica da literatura foi realizada sistematicamente, a fim de identificar trabalhos propondo técnicas similares a BPPA, que incluíssem alguma validação. Como resultado, poucos trabalhos com esse objetivo foram encontrados, e os encontrados não tem foco específico em BPM. Para fins da análise de literatura efetuada, questões de pesquisa como "Quais as medidas para medição de tamanho funcional de BP encontradas?"; "Quais os métodos de validação destas medidas?"; e "Há trabalhos que tratem especificamente sobre medidas para BPM com FPA?"foram respondidas com a utilização de 20 estudos identificados relevantes para tal, onde destes, 23 técnicas de medição foram encontradas, dentre elas 11 focadas na medição funcional em geral, não aplicadas a BPM, e apenas sete apresentam algum processo de validação conforme esta proposta de pesquisa. A análise da bibliografia realizada permitiu definir com mais clareza a justificativa de realização deste projeto de pesquisa, após as lacunas na bibliogafia terem sido identificadas.

Com o objetivo de definir maiores detalhes relacionados ao experimento a ser realizado, informações adicionais estão sendo coletadas e reunião com especialistas nesse tipo de projeto estão sendo realizadas. O auxílio de especialistas da área é de suma relevância para o projeto de pesquisa, pois permite maior obter maior conhecimento relacionado ao cenárop de aplicação, direcionando assim, esforços de pesquisa com maior precisão. Além disso, um estudo cuidadoso de possíveis técnicas estatísticas que auxiliem a análise e compartilhamento dos dados e resultados do estudo deve ser iniciado em breve, o qual deverá influenciar no detalhamento do experimento permitindo melhor apresentar os dados quantitativos da proposta.

O exame de qualificação será realizado entre junho e agosto deste ano, e o texto a ser apresentado na qualificação está em elaboração. Após o exame de qualificação, o experimento será refinado considerando as possíveis sugestões da banca. O experimento deverá ser realizado no final do segundo semestre de 2017, de forma que os dados obtidos sejam avaliados no início de 2018. A escrita da dissertação deve ocorrer entre o final de 2017 e início de 2018, culminando no depósito da dissertação em março de defesa em abril de 2018.

5. CONCLUSÃO

Este projeto de pesquisa busca realizar um experimento que possa confirmar os benefícios da técnica BPPA e apresentar suas possíveis limitações. Considerando a potencial importância de BPPA para gerentes de projetos de BPM, é importante que eles possam confiar na aplicação da técnica, por meio de evidência a serem apresentadas como resultado deste projeto. Para isso, um experimento sistemático está em planejamento, para que seus resultados possam ser considerados consistentes e com poucas ameaças à validade. O planejamento do experimento ainda precisa ser finalizado para que seja realizado no segundo semestre deste ano. Espera-se ao final que outros pesquisadores possam propor melhorias para BPPA ou então novas técnicas para o mesmo objetivo que possam ser diretamente comparadas com BPPA.

[1]'Corretude' está sendo usado, em vez do termo estatístico 'validade', para diferenciar da 'validação' geral pretendida.

6. REFERÊNCIAS

[1] S. Abrahão, G. Poels, and O. Pastor. A functional size measurement method for object-oriented conceptual schemas: design and evaluation issues. *Software & Systems Modeling*, 5(1):48–71, 2006.

[2] A. Arsanjani, N. Bharade, M. Borgenstrand, P. Schume, J. K. Wood, V. Zheltonogov, et al. *Business Process Management Design Guide: Using IBM Business Process Manager*. IBM Redbooks, 2015.

[3] M. Baklizky, M. Fantinato, L. H. Thom, V. Sun, E. P. Prado, and P. Hung. Business process points-a proposal to measure bpm projects. In *Proc. of the 21st European Confer. on Information Systems (ECIS)*, page 2, 2013.

[4] M. Fantinato, I. M. S. Gimenes, and M. B. F. Toledo. *Product line in the business process management domain*, pages 497–530. Boston, MA, USA, Auerbach Publications, 2010.

[5] R. Feldt and A. Magazinius. Validity threats in empirical software engineering research - an initial survey. In *Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering (SEKE)*, pages 374–379, 2010.

[6] D. Longstreet. Fundamentals of function point analysis. *Longstreet Consulting, Inc*, 2002.

[7] G. H. Travassos, D. Gurov, and E. A. G. do Amaral. Introdução à engenharia de software experimental. Technical report, PESC, COPPE, UFRJ, Brasil, 2002.

[8] T. Uemura, S. Kusumoto, and K. Inoue. Function-point analysis using design specifications based on the unified modelling language. *J. of software maintenance and evolution: Research and practice*, 13(4):223–243, 2001.

[9] M. Weske. Business process management architectures. In *Business Process Management*, pages 333–371. Springer, 2012.

Organização:

UFLA
UNIVERSIDADE FEDERAL DE LAVRAS

Afiliado à:

Cooperação:

acm

Fomento:

CAPES 65 anos

Patrocínio Prata:

SISTEMA INFORMÁTICA

Apoio:

compJÚNIOR

CENTRO ACADÊMICO DE SISTEMAS DE INFORMAÇÃO

INNERVISION

30 ANOS SQUADRA Tecnologia