Embeddings para previsão da próxima atividade em Mineração de Processos

Nicolas Jashchenko Omori[1], **Sylvio Barbon Jr.**[1]

[1]Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – CEP 86057-970 – Londrina – PR – Brasil

omori@uel.br, barbon@uel.br

Abstract. *Each day, the use of computational systems is growing. Knowing this, organizations are left with a massive amount of data that describes their processes. Therefore, a demand arises on the extraction of important insights of this data, aiming for a competitive advantage over other organizations. The prediction on the next activity can optimize the flow of a real life process and, to make that prediction will be used embedding and machine learning techniques.*

Resumo. *Com o aumento a cada dia no uso de sistemas computacionais, organizações contam com uma enorme quantidade de dados que descrevem seus processos. Com isso, surge uma demanda para a extração de informações importantes desses dados, a fim de obter uma vantagem competitiva em relação a outras organizações. A predição da próxima atividade pode otimizar o andamento de um processo na vida real e, para realizar essa predição, serão utilizadas técnicas de embedding e aprendizado de máquina.*

1. Introdução

Empresas e organizações estão, a cada dia, mais dependentes de sistemas computacionais para a realização de seus processos, tanto internos quanto externos, assim gerando uma grande quantidade de dados sobre estes. Com esse grande crescimento na quantidade de dados, surge o interesse em extrair informações importantes para as organizações, a fim de compreender e aprimorar seus processos. Uma das formas de se obter uma vantagem competitiva em seus processos é a predição da próxima atividade em uma certa instância de um processo, possibilitando a alocação preemptiva de recursos para a realização da atividade prevista.

Visando esse interesse, surge a Mineração de Processos, uma área multidisciplinar que utiliza técnicas de Mineração de Dados, Business Intelligence (BI) e Business Process Management (BPM), visando extrair informações factuais dos dados de processos. Apesar de usar técnicas das disciplinas citadas, a Mineração de Processos não é apenas a junção dessas técnicas, mas também uma aplicação mais centrada nos processos, visto que a Mineração de Dados é focada demais nos dados para fazer uma análise de um processo do começo ao fim, BI foca apenas em *dashboards* e reportes ao invés de *insights* precisos, enquanto técnicas de BPM são aplicadas em processos idealizados por especialistas, que normalmente são diferentes dos processos reais de uma organização [11].

O restante do documento está dividido da seguinte forma: A Seção 2 apresenta os conceitos fundamentais para a realização do trabalho, assim como trabalhos relacionados; a Seção 3 apresenta os objetivos a serem alcançados; a Seção 4 define os métodos e

técnicas que serão utilizados para atingir os objetivos; a Seção 5 define e mostra um cronograma das atividades até o final do projeto; e, por fim, na Seção 6 são apresentadas as contribuições e resultados esperados após a conclusão do projeto.

2. Fundamentação Teórico-Metodológica e Estado da Arte

Nas seguintes subseções serão apresentados os conceitos fundamentais para a compreensão e realização deste trabalho, assim como a pontuar trabalhos relacionados em predição da próxima atividade e representação de processos como vetores de características.

2.1. Process Mining

Process Mining (PM) é uma área de processos que tem como objetivo descobrir, checar conformidade e aprimorar *Business Processes (BP)*, sendo dividida entre essas três ações (em inglês, *Discovery, Conformance Checking e Enhancement*) [11].

A área de *Discovery* tem como objetivo extrair um modelo de um *log* de evento, a área de *Conformance Checking* verifica se um modelo representa de forma apropriada o processo da vida real e a área de *Enhancement* propõe maneiras de aprimorar o processo utilizando as informações extraídas do modelo e do *log* de evento. A imagem 1 mostra como as três áreas se relacionam entre si e com o mundo externo.

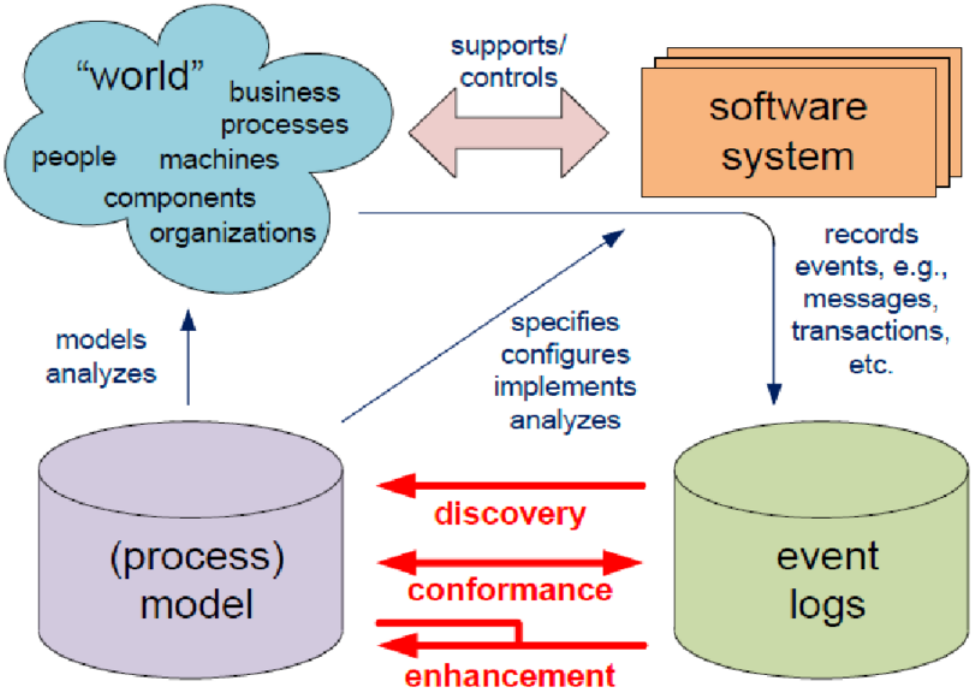

Figura 1. Área de Process Mining, retirada de [11]

2.2. Conceitos em Process Mining

Os dados de um evento são comumente armazenados em Event logs, que possuem as seguintes definições [11]:

Definição 1 (Eventos, atributos, *traces*): Seja Σ o universo de eventos, ou seja, o conjunto de todos os possíveis identificadores de evento. Eventos podem possuir diversos *atributos*, como um *timestamp*, a atividade, o recurso, o custo, entre outros. Seja AN o conjunto de nomes de atributos. Para cada evento $e \in \Sigma$ e nome $n \in AN$, então $\#_n(e)$ é o valor do atributo n para o evento e, se o evento e possuir um atributo n, senão $\#_n(e)$ é nulo. Um *trace* é uma sequência não vazia de eventos $\sigma \in \Sigma^*$ onde cada evento aparece uma única vez e o tempo é não decrescente, ou seja, for $1 \leq i < j \leq |\sigma|$: $\sigma(i) \neq \sigma(j)$.

Definição 2 (*Cases*, *Log* de Evento): Seja $\mathcal{C}$ o universo de *cases*, ou seja, o conjunto de todos os possíveis identificadores de cases. Um *Event Log* é um conjunto de cases $L \subseteq \mathcal{C}$ onde cada evento aparece apenas uma vez no *log*, ou seja, para dois diferentes cases, a intersecção de seus conjuntos de eventos deve ser vazia.

De forma mais simples, um *log* de evento possui diferentes *cases*, que possuem um *trace*, que são uma sequência de eventos com *timestamps* não-decrescentes. Na figura 2, temos um exemplo de *log* de evento, retirado de [11]. Neste *log*, cada evento possui os atributos Activity id, Originator e Timestamp e estão inseridos em diversos cases. Um exemplo de *trace* nesse log seria $< activityA, activityB, activityC, activityD >$, presente no *case 1*.

Case id	Activity id	Originator	Timestamp
case 1	activity A	John	9-3-2004:15.01
case 2	activity A	John	9-3-2004:15.12
case 3	activity A	Sue	9-3-2004:16.03
case 3	activity B	Carol	9-3-2004:16.07
case 1	activity B	Mike	9-3-2004:18.25
case 1	activity C	John	10-3-2004:9.23
case 2	activity C	Mike	10-3-2004:10.34
case 4	activity A	Sue	10-3-2004:10.35
case 2	activity B	John	10-3-2004:12.34
case 2	activity D	Pete	10-3-2004:12.50
case 5	activity A	Sue	10-3-2004:13.05
case 4	activity C	Carol	11-3-2004:10.12
case 1	activity D	Pete	11-3-2004:10.14
case 3	activity C	Sue	11-3-2004:10.44
case 3	activity D	Pete	11-3-2004:11.03
case 4	activity B	Sue	14-3-2004:11.18
case 5	activity E	Clare	17-3-2004:12.22
case 5	activity D	Clare	18-3-2004:14.34
case 4	activity D	Pete	19-3-2004:15.56

Figura 2. *Log* de evento retirado de [11]

2.3. Predição em Process Mining

Em [8] foi feito um levantamento das técnicas disponíveis para predição em *Business Processes*, no qual foram analisadas 41 publicações, que foram separadas em relação às

técnicas utilizadas e o tipo de informação prevista por essas propostas.

Nesse trabalho, são levantados oito métodos para a predição da próxima atividade, com as técnicas utilizadas abrangendo desde métodos estatísticos [1, 2, 5], técnicas de aprendizado de máquina supervisionadas (e.g. árvores de decisão, SVM) [13, 10, 4] e não supervisionadas (e.g. *clustering* e redes neurais) [12, 9].

2.4. Aprendizado de Máquina

Aprendizado de Máquina é uma área de conhecimento preocupada em escrever programas de computadores que se melhoram automaticamente para uma certa tarefa. Mais formalmente, Mitchell [7] define: "Um programa de computador aprende com uma experiência E, com respeito a uma classe de tarefas T e uma medida de performance P, se sua performance nas tarefas em T, medidas por P, é melhor com a experiência E".

Exemplificando, experiência E pode ser um conjunto de dados, como uma tabela com a área total e o preço de casas, a tarefa T pode ser a predição do preço de uma casa dada sua área total e a medida de performance pode ser o erro médio na predição de preço dessas casas.

As técnicas de aprendizado de máquina são separadas entre supervisionada ou não-supervisionada, a primeira possuindo as *respostas* do problema em sua fase de treino e a segunda não possuindo.

2.5. Aprendizagem de Representação em Process Mining

Act2Vec, Trace2Vec, Log2Vec e Model2Vec são técnicas propostas por De Koninck et al.[3] e são utilizadas para a representação de atividades, *traces*, *logs* e modelos de processo como vetores de características para serem usados com técnicas de Aprendizado de Máquina.

As técnicas utilizam conceitos de mineração de texto, considerando uma atividade como uma palavra, um *trace* como uma sentença e um *log* como um documento, e utilizam a técnica Word2Vec [6], que é amplamente utilizada na área de mineração de texto.

A publicação [3] é o primeiro trabalho a respeito da utilização de técnicas de *Embedding* de palavras e Aprendizado de Representação para a resolução de problemas na área de Process Mining.

3. Objetivos

O objetivo principal do trabalho é propor uma técnica para a predição da próxima atividade de um *case* em um processo, utilizando conceitos de mineração de texto e aprendizagem de máquina como base da técnica proposta e, a partir disso, obter resultados comparáveis ou superiores a técnicas propostas no estado da arte.

4. Procedimentos metodológicos/Métodos e técnicas

Inicialmente, será feita uma revisão bibliográfica das técnicas e conceitos a serem utilizados ao decorrer do trabalho, levantando trabalhos relacionados para a análise das propostas do estado da arte.

Depois disso, será feito o levantamento dos logs de evento a serem utilizados no projeto, priorizando, mas não se restringindo a, bases comumente utilizadas por trabalhos correlatos para a comparação de resultados com as técnicas atuais.

O projeto então será implementado e avaliado, comparando os resultados da proposta com o estado da arte. Por fim, será feita a escrita do documento final.

5. Cronograma de Execução

Atividades:

1. Estudo das técnicas e conceitos a serem utilizados;
2. Levantamento de logs de evento;
3. Implementação do projeto;
4. Avaliação dos resultados;
5. Escrita do documento;

Tabela 1. Cronograma de Execução

	abr	mai	jun	jul	ago	set	out	nov
Atividade 1	x	x	x					
Atividade 2		x	x					
Atividade 3			x	x	x	x		
Atividade 4					x	x	x	
Atividade 5					x	x	x	x

6. Contribuições e/ou Resultados esperados

Com esse trabalho, é esperado que organizações consigam obter uma vantagem competitiva ao utilizar as técnicas propostas de predição de próxima atividade. Com um método de predição eficiente, é possível alocar de forma preemptiva os recursos necessários para a execução de uma certa atividade em uma instância de um processo.

7. Espaço para assinaturas

Londrina, 08 de abril de 2019.

Aluno

Orientador

Referências

[1] Jörg Becker, Dominic Breuker, Patrick Delfmann, and Martin Matzner. Designing and implementing a framework for event-based predictive modelling of business processes. In Fernand Feltz, Bela Mutschler, and Benoît Otjacques, editors, *Enterprise modelling and information systems architectures - EMISA 2014*, pages 71–84, Bonn, 2014. Gesellschaft für Informatik e.V.

[2] Dominic Breuker, Patrick Delfmann, Martin Matzner, and Jörg Becker. Designing and evaluating an interpretable predictive modeling technique for business processes. In *Business Process Management Workshops*, 2014.

[3] Pieter De Koninck, Seppe vanden Broucke, and Jochen De Weerdt. act2vec, trace2vec, log2vec, and model2vec: Representation learning for business processes. In Mathias Weske, Marco Montali, Ingo Weber, and Jan vom Brocke, editors, *Business Process Management*, pages 305–321, Cham, 2018. Springer International Publishing.

[4] Patrick Delfmann Dominic Breuker, Martin Matzner and Jörg Becker. Comprehensible predictive models for business processes. 40:1009–1034, 2016.

[5] Geetika T. Lakshmanan, Davood Shamsi, Yurdaer N. Doganata, Merve Unuvar, and Rania Khalaf. A markov prediction model for data-driven semi-structured business processes. *Knowledge and Information Systems*, 42:97–126, 2013.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR (Workshop)*, 2013.

[7] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[8] A. E. Márquez-Chamorro, M. Resinas, and A. Ruiz-Cortés. Predictive monitoring of business processes: A survey. *IEEE Transactions on Services Computing*, 11(6):962–977, Nov 2018.

[9] Niek Tax, Ilya Verenich, Marcello La Rosa, and Marlon Dumas. Predictive business process monitoring with lstm neural networks. In Eric Dubois and Klaus Pohl, editors, *Advanced Information Systems Engineering*, pages 477–492, Cham, 2017. Springer International Publishing.

[10] Merve Unuvar, Geetika T. Lakshmanan, and Yurdaer N. Doganata. Leveraging path information to generate predictions for parallel business processes. 47(2):433–461, 2016. Exported from https://app.dimensions.ai on 2019/03/28.

[11] Wil M.P. van der Aalst. *Process Mining Discovery, Conformance and Enhancement of Business Processes*. Springer-Verlag Berlin Heidelberg, 2011.

[12] Ilya Verenich, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, and Chiara Di Francescomarino. Complex symbolic sequence clustering and multiple classifiers for predictive process monitoring. In Manfred Reichert and Hajo A. Reijers, editors, *Business Process Management Workshops*, pages 218–229, Cham, 2016. Springer International Publishing.

[13] Ilya Verenich, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, and Chiara Di Francescomarino. Minimizing overprocessing waste in business processes via predictive activity ordering. In *28th International Conference on Advanced Information Systems Engineering (CAiSE2016)*, pages 186–202, Ljubljana, Slovenia, 2016. Proceedings published as: Advanced Information Systems Engineering; Volume 9694 of the series Lecture Notes in Computer Science.