Análise de Colaboração em Processos de Negócio por meio de SGBDs de Grafos e Dados de Proveniência Multimodais*

Maria Luiza Falci[1], **Andréa Magalhães**[1], **Vanessa Braganholo**[1], **Aline Paes**[1], **Daniel de Oliveira**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
CEP – 24210-346 – Niterói – RJ – Brasil

marialuizafalci@id.uff.br

{andrea, vanessa, alinepaes, danielcmo}@ic.uff.br

Resumo. *Durante a definição e execução de processos de negócio, dados de proveniência em diferentes formatos são coletados, e analisá-los de forma integrada é uma tarefa complexa e propensa a erros, se realizada de forma manual. Entretanto, tal integração pode trazer insights sobre o processo. O presente artigo apresenta a abordagem MINERVA (Multimodal busINEss pRoVenance Analysis), que permite a análise de colaboração e identificação de pontos de melhoria em processos de negócio por meio de dados de proveniência multimodais e Bancos de Dados orientados a grafos. A abordagem foi avaliada por meio de um estudo de viabilidade com dados reais de uma empresa de consultoria.*

Abstract. *Data provenance in different formats are collected throughout the execution and definition of business processes. Analyzing these data in an integrated form can be a complex and error prone task when performed manually. However, this analysis can generate insights about the business process. This work presents MINERVA (Multimodal busINEss pRoVenance Analysis), an approach that focuses on collaboration analysis and on identifying possible improvements in the business process using multimodal data provenance and Graph Databases. The proposed approach was evaluated through a feasibility study that used real data from a consulting company.*

1. Introdução

Atualmente, o trabalho nas organizações possui mais colaboração com o objetivo de otimizar a execução de tarefas complexas com mais qualidade [Mathiesen et al. 2012]. Analisar e compreender como as pessoas colaboram dentro da organização se torna importante devido ao crescimento de equipes multidisciplinares. As organizações comumente estruturam seus processos no formato de modelos de processos de negócio, utilizando abordagens e ferramentas que oferecem apoio à modelagem, execução e melhoria dos processos [Dumas et al. 2013]. Além disso, de forma a se alcançar a melhoria dos processos, é fundamental que existam medidas de desempenho para tais processos (*e.g.*, tempo de execução).

Uma vez definidas as medidas, pode-se realizar uma análise da execução do processo para verificar se existem pontos de melhoria. E para realizar essas análises, dados

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq e FAPERJ.

históricos se fazem necessários. Esses dados devem possuir semântica e serem passíveis de consulta. Dessa forma, dados de proveniência [Freire et al. 2008] podem ser aplicados nesse cenário. Os dados de proveniência descrevem a origem de um dado, assim como o processo pelo qual ele passou até seu estado final. Os dados de proveniência permitem que, ao encontrar algum fragmento de dado que não tenha um comportamento esperado, se possa buscar a sua origem e as transformações que foram feitas até o momento atual. No contexto da análise de processos de negócio, os dados de proveniência permitem identificar quais colaboradores contribuíram para uma determinada atividade ou processo. Além disso, os dados de proveniência informam as fontes dos dados utilizadas nos processos, quais atividades foram atribuídas a quais colaboradores, *etc*.

Entretanto, os dados de proveniência de processos de negócio são naturalmente multimodais, *i.e.*, podem ser representados de diversas formas (*e.g., logs* de eventos, textos, áudio, *etc*.). Todas estas informações devem ser levadas em consideração no momento da análise de colaboração e conformidade para checar se a execução do processo está de acordo com seu modelo, ou no momento de prover suporte operacional, *i.e.,* apoio à execução enquanto ela ocorre [van der Aalst 2016]. A análise de dados de proveniência multimodais oferece possibilidades de se obter novos conhecimentos por meio da integração de uma ampla variedade de dados provenientes de fontes heterogêneas. A maioria dos trabalhos encontrados na literatura que analisam a colaboração em processos de negócio utiliza somente *logs* de eventos como dados de entrada, e não conseguem lidar com dados em formatos heterogêneos. Van Der Aalst *et al.* [2005] e Van Der Aalst [2016] descobrem redes sociais a partir de *logs* de eventos, e conseguem capturar apenas os nomes a quem uma atividade foi delegada. Ferreira e Alves [2011] também extraem redes sociais de *logs*, e afirmam que as redes sociais extraídas podem ser muito complexas. Zhao e Zhao [2014] apresentam uma visão geral sobre trabalhos que utilizam *logs* de eventos para extração da estrutura organizacional, rede social, funções de atores dos processos e alocação de recursos.

Capturar e analisar dados de proveniência multimodais no contexto de processos de negócio é um problema em aberto, e também um desafio. Ainda que existam padrões de representação de proveniência como o W3C PROV [Belhajjame et al. 2012], que pode ser estendido e aplicado a diferentes contextos, eles não tratam da captura e processamento de dados multimodais, se limitando a representar a proveniência de forma estruturada. Assim, com o objetivo de analisar dados de proveniência multimodais, a colaboração e identificar pontos de melhoria no processo do negócio, o presente artigo propõe uma abordagem chamada MINERVA (*Multimodal busINEss pRoVenance Analysis*). Em sua versão atual, a MINERVA é capaz de processar dados de *logs*, documentos em texto livre e áudio. A MINERVA foi avaliada com dados reais da empresa dheka Consultoria[1] e os resultados se mostraram promissores. O presente artigo está dividido como segue. A Seção 2 apresenta a arquitetura proposta da abordagem MINERVA. Na Seção 3 é descrito um estudo de viabilidade. Na Seção 4, apresentamos as considerações finais e trabalhos futuros.

2. Abordagem Proposta: MINERVA

Dados sobre processos de negócio podem ser analisados com diferentes objetivos. Tradicionalmente, a análise desses dados utiliza como entrada *logs* de eventos. Entretanto, a

[1] `https://www.dheka.com.br/`

execução de processos de negócio pode estar relacionada a diferentes dados, como *e-mails*, ou dados de plataformas que dão apoio à execução das atividades. Esses dados representam a proveniência do processo e devem ser considerados no momento da análise. A arquitetura da abordagem proposta é apresentada na Figura 1, e é composta por 3 camadas: (i) Fontes de Dados Externas, (ii) Camada de Processamento, e (iii) Camada de Dados. As Fontes de Dados Externas representam os dados de proveniência do processo que podem ser representados em múltiplos formatos. Em sua versão atual, a MINERVA considera textos de *e-mails*, conversas de aplicativos de mensagem instantânea (*e.g.*, WhatsApp), vídeos de reuniões e comentários em texto livre, que podem ser escritos em plataformas de apoio à execução das atividades do processo (*e.g.*, Pipedrive). Tais comentários podem conter dados de proveniência importantes e que definem como se dá a colaboração durante a execução do processo.

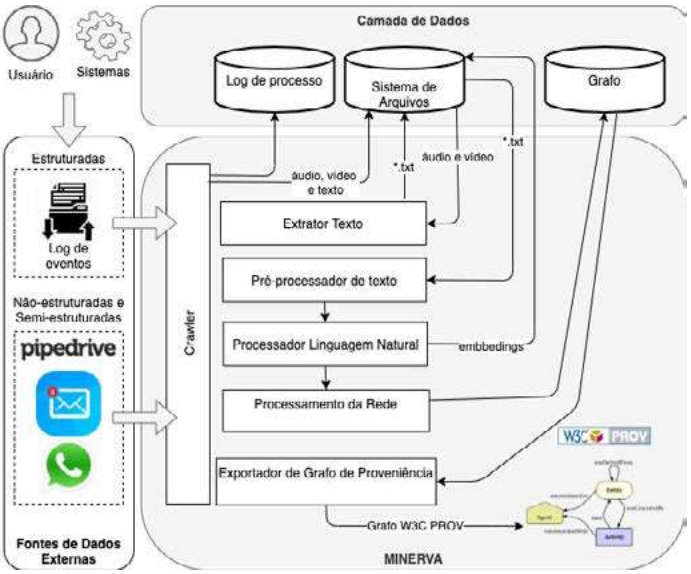

Figura 1. Arquitetura da Abordagem MINERVA

Na Camada de Processamento todo o conteúdo das fontes de dados externas é capturado pelo componente *Crawler* e armazenado na Camada de Dados (explicada a seguir). O *Crawler* é um agente que recebe como entrada uma lista de endereços de recursos para visitar (*e.g., seeds*). À medida que o *Crawler* visita esses endereços, ele identifica o dado de proveniência (seja *log*, texto ou vídeo) e o insere no banco de dados de proveniência ou no sistema de arquivos (dependendo do seu tipo). Como os dados de proveniência extraídos podem ser vídeo ou áudio, o *Extrator de Texto* converte o conteúdo para texto. Tais dados são armazenados em um diretório do sistema de arquivos e referenciados no BD de proveniência (somente o caminho do arquivo bruto). Uma vez que todos os dados não-estruturados se encontram em formato textual, o *Processador de Linguagem Natural* identifica no texto elementos como substantivos, verbos, *etc*, que são usados para identificar o histórico de ações e a rede de colaboração. Esse componente pode ser implementado de múltiplas formas, mas na versão atual usa o Spacy [Honnibal and Montani 2017], que é uma ferramenta de processamento de linguagem natural. Com o Spacy podemos fazer normalizações no texto, *e.g.*, lematização ou *stemming*, remover *stopwords* e remover algumas categorias, com POS (*Part of Speech*), *etc*. Os textos são gravados também em sua forma vetorial (*i.e., embbedings*), de forma que o processamento por outros algoritmos no

futuro (*e.g.*, identificação de tópicos) seja facilitado.

Finalmente, de posse de informações de substantivos e verbos, o grafo de colaboração é criado no componente de *Processamento da Rede*. Os nomes coletados pelo *Processador de Linguagem Natural* são utilizados para criar nós de uma rede complexa, e as arestas são criadas de acordo com regras pré definidas (*e.g.*, na frase "Maria ligou para João", seriam criados dois nós [Maria e João], e seria criada uma aresta entre estes nós, que teria o peso influenciado pelo verbo "ligou"). Os tipos possíveis de arestas são definidos de acordo com o modelo de proveniência apresentado na Figura 2, que segue os relacionamentos expressos no padrão PROV para representar os dados do estudo de caso apresentado na Seção 3. A rede complexa criada representa uma rede social dos participantes das execuções dos processos. Muitos nomes que aparecem nessa rede provavelmente não são vistos nos *logs* de eventos, já que apesar de uma atividade ser delegada a uma pessoa, ela pode colaborar com outras para executá-la (*e.g.*, ao pedir ajuda). Além de gerar o grafo de colaboração, o componente de *Processamento da Rede* também analisa o papel de cada nó da rede (que representa pessoas ou organizações), e detalhes como centralidade, grau de entrada e grau de saída. Com essas análises, é possível compreender melhor a influência que cada pessoa teve sobre a execução de um determinado processo ou atividade [Cross et al. 2004]. Finalmente, o componente *Exportador de Grafo de Proveniência* gera um arquivo *JSON* contendo o grafo gerado de acordo com o modelo supracitado. Tal grafo é validado usando serviços como o ProvToolBox[2]. Finalmente, a Camada de Dados contém os dados de proveniência (atualmente em um banco de dados relacional PostgreSQL), o sistema de arquivos que armazena os dados brutos de áudio, vídeo, texto e *embeddings* e o BD de grafos que armazena a rede de colaboração (atualmente utilizamos o Neo4J).

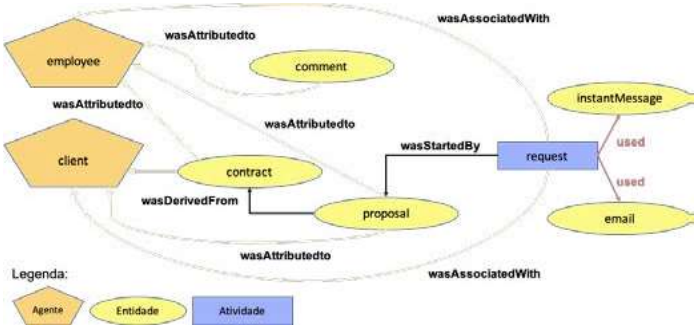

Figura 2. Modelo de Proveniência

3. Avaliação Experimental

Para avaliar a MINERVA, realizamos um estudo de caso com dados reais de uma empresa de consultoria. Os dados utilizados são sobre o processo comercial da empresa dheka, oriundos de *logs* de eventos e do Pipedrive (que contém comentários extraídos de e-mails e trocas de mensagens). No contexto do estudo de caso (plataforma de CRM), cada venda em potencial (que pode ser concretizada ou não), é considerada como uma instância do processo comercial. Cada instância possui dados sobre as atividades do processo comercial que foi executado, qual colaborador ou cliente está relacionado a atividade, dados sobre o

[2] https://lucmoreau.github.io/ProvToolbox/

cliente, além de comentários em texto livre. O modelo de proveniência de dados do processo de vendas é apresentado na Figura 2. Podemos observar que a origem dos dados se dá a partir de uma requisição (*request*), que está relacionada a dados de *e-mail* e mensagens instantâneas (*instantMessage*). A partir de cada requisição (*request*), uma proposta de negócio (*proposal*) é gerada, que possui relação com o cliente (*client*) e o colaborador da empresa (*employee*), e pode se tornar um contrato (*contract*) (após revisada e assinada). O colaborador da empresa também está relacionado a comentários (*comment*), que ele pode escrever livremente sobre o processo. Tradicionalmente, a análise dos dados sobre as execuções de atividades do processo comercial usam somente *logs* de eventos. Entretanto, como o foco da abordagem desse artigo é na análise de dados que tragam informações adicionais, os comentários em texto livre foram escolhidos para análise.

Foi realizado o *download* dos comentários diretamente do Pipedrive no formato *CSV*. Cada comentário pode conter mais de uma frase. Os comentários possuem diversas palavras abreviadas e siglas, que foram adicionadas a um dicionário, que foi utilizado para substituir essas palavras para sua representação não-abreviada. Após a transformação das palavras, cada comentário serviu como entrada para o Spacy [Honnibal and Montani 2017] em português, que extraiu os nomes e verbos de cada frase. Para criar a rede, foi utilizado o Neo4j. Os nós, arestas e consultas sobre a rede complexa foram criados utilizando a linguagem Cypher. Os nomes das pessoas e organizações foram utilizados para criar os nós da rede. As arestas foram criadas seguindo regras pré-definidas que buscam identificar quando existe um relacionamento entre duas pessoas (*e.g.*, quando uma contactou a outra). Os verbos foram utilizados para definir os pesos das arestas. Após a criação do BD, foi possível observar na rede todas as pessoas e organizações que influenciaram nas execuções do processo, conforme apresentado na Figura 3. Com a construção da rede de colaboração completa (fragmento apresentado na Figura 3) a partir dos comentários sobre o processo, foi possível adicionar pessoas que colaboraram com a execução dos processos e não apareciam oficialmente no processo. Oficialmente cada instância do processo estava relacionada apenas à uma pessoa e, como representado na Figura 3, em alguns casos (como na instância de processo “117”) foram identificadas 4 pessoas relacionadas. A partir dessas análises é possível responder questões sobre como é de fato realizada a colaboração no processo, *e.g.,* quais pessoas trabalham em conjunto, quem pode ajudar a resolver alguma questão do processo, *etc*.

4. Considerações Finais

A abordagem proposta nesse artigo, denominada MINERVA, foca em analisar dados de execuções de processos de negócio, se diferenciando das técnicas usuais da literatura em múltiplos aspectos. O primeiro ponto de diferença são os dados de proveniência utilizados, que podem ser heterogêneos, não se limitando a *logs* de eventos. Os dados utilizados no estudo de caso foram dados textuais originados de comentários textuais, além dos *logs* de eventos. O segundo ponto de diferença é o foco na rede de colaboração oriunda dos comentários textuais, rede essa que era “invisível” nos dados de proveniência encontrados nos *logs* de eventos. Essa rede social permitiu que análises complementares pudessem ser realizadas sobre uma rede que se encontrava oculta inicialmente. Outros pontos importantes da abordagem foram o modelo de proveniência proposto (compatível com o W3C PROV) e a criação da rede social utilizando um BD orientado à grafos, abordagem que até o presente momento não foi proposta em nenhum dos trabalhos relacionados. O modelo

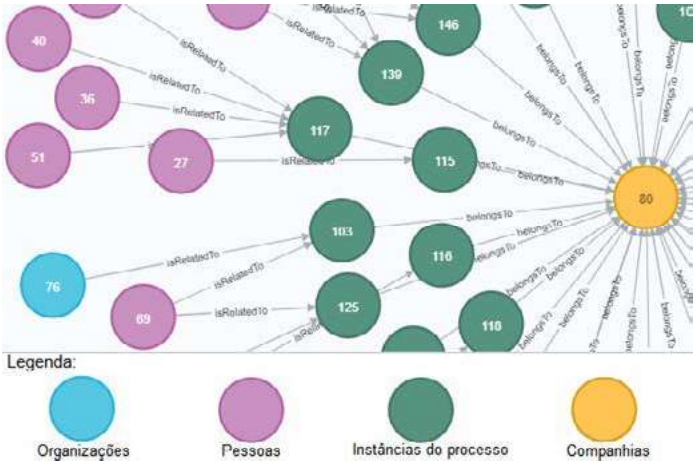

Figura 3. Fragmento da Rede de Colaboração do Estudo de Caso

de proveniência permite que a origem dos dados possa ser investigada, além de todo o processo de transformação dos dados (*e.g.*, qual contrato foi gerado a partir de qual proposta). Adicionalmente, o BD em grafo permite que os dados sejam analisados sob o ponto de vista de uma rede complexa. Em trabalhos futuros, novos estudos de caso serão realizados com diferentes fontes de dados, como mensagens de texto instantâneas, gravações de reuniões e *e-mails*. Além disso, serão apresentadas análises sobre a rede complexa.

Referências

Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., and Zednik, S. (2012). Prov model primer. *World Wide Web Consortium*.

Cross, R. L., Cross, R. L., and Parker, A. (2004). *The hidden power of social networks: Understanding how work really gets done in organizations*. Harvard Business Press.

Dumas, M., Rosa, M. L., Mendling, J., and Reijers, H. A. (2013). *Fundamentals of Business Process Management*. Springer.

Ferreira, D. R. and Alves, C. (2011). Discovering user communities in large event logs. In *International Conference on Business Process Management*, pages 123–134. Springer.

Freire, J., Koop, D., Santos, E., and Silva, C. T. (2008). Provenance for computational tasks: A survey. *Comput. Sci. Eng.*, 10(3):11–21.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Mathiesen, P., Watson, J., Bandara, W., and Rosemann, M. (2012). Applying social technology to business process lifecycle management. In *BPM Workshops*, pages 231–241. Springer.

van der Aalst, W. (2016). *Process Mining: Data Science in Action*. Springer Publishing Company, Incorporated, 2nd edition.

Van Der Aalst, W. M., Reijers, H. A., and Song, M. (2005). Discovering social networks from event logs. *Computer Supported Cooperative Work (CSCW)*, 14(6):549–593.

Zhao, W. and Zhao, X. (2014). Process mining from the organizational perspective. In *Foundations of intelligent systems*, pages 701–708. Springer.