Mineração de Textos usando Agrupamento Hierárquico e Reconhecedor de Entidades Nomeadas em um ambiente de investigação policial*

Fabrício J. Barth, Maria Cristina Belderrain,
Nádia L. P. Quadros, Luciane L. Ferreira, Antonio P. Timoszczuk

[1]Diretoria de Inovações e Soluções Tecnológicas
Fundação Atech Tecnologias Críticas (http://www.atech.br)

{fbarth, mbelderrain, nquadros, lferreira, antoniop} @atech.br

Resumo. *Este trabalho descreve um processo de mineração de texto composto por algoritmos de agrupamento e de identificação de entidades nomeadas. Durante a avaliação do trabalho verificou-se que a combinação dos resultados dos algoritmos sob a forma de gráficos agrega valor à investigação policial. A apresentação gráfica permite que o investigador construa cenários visuais de investigações em andamento e, de forma interativa, explore os dados e monitore eventuais mudanças nos cenários. O acréscimo de rótulos aos agrupamentos contribuiu significativamente para a melhor interpretação dos resultados.*

1. Introdução

Situações complexas da investigação criminal exigem um processo de transformação de grandes volumes de dados díspares em informações sintéticas e conclusivas [Júnior and de Lima Dantas 2006]. Tipicamente, as fontes de informação utilizadas por investigadores são: (i) notícias de veículos jornalísticos, (ii) conteúdo de boletins de ocorrência extraídos a partir de um banco de dados, (iii) textos gerados a partir de escutas telefônicas, (iv) informações sobre inquéritos policiais extraídos a partir de um banco de dados ou relatórios em formato texto e (v) informações sobre denúncias extraídas a partir de um banco de dados ou relatórios em formato texto. Na maioria dos casos, o investigador está interessado em verificar a existência de elementos associados, identificar relações entre fatos e construir modelos de informação sintetizada sobre a investigação.

Este trabalho descreve em detalhes o processo de mineração de texto utilizado em [Barth et al. 2007], onde as fontes de informação são submetidas a algoritmos de agrupamento e de identificação de entidades nomeadas (pessoas, organizações, locais e termos importantes para o domínio da aplicação). Ambas são técnicas de mineração de texto que colocam em evidência as relações entre as entidades e permitem a apresentação gráfica das mesmas ao usuário, figura 1. A avaliação realizada neste trabalho é uma continuação da avaliação realizada em [Barth et al. 2007]. A criação de rótulos para os nodos no agrupamento hierárquico foi uma necessidade identificada em [Barth et al. 2007] e implementada neste trabalho.

Este texto está estruturado da seguinte maneira: na seção 2 é apresentado o método para determinação de similaridades entre os documentos; na seção 3 é apresentado o

*Este trabalho foi parcialmente financiado pelo CNPq através de duas bolsas RHAE, processo número 520092/06-6.

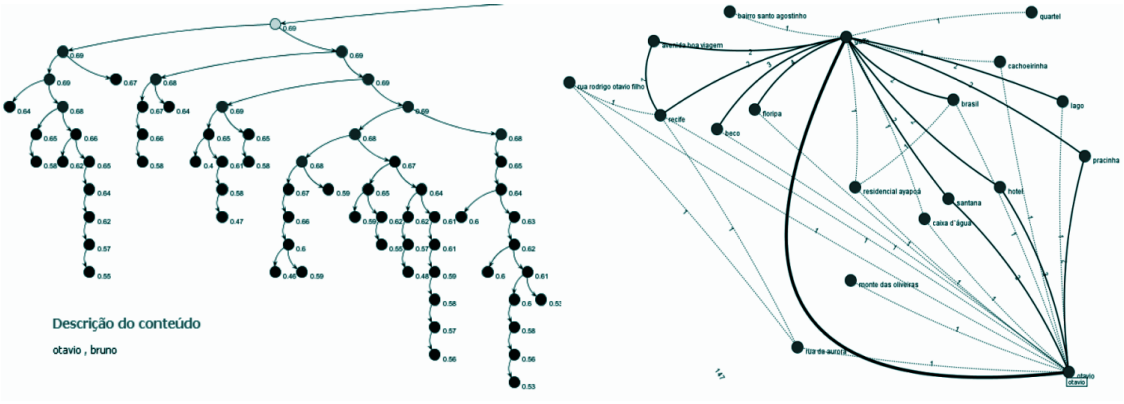

Figura 1. Exemplo de (a) agrupamento hierárquico e (b) grafo de relacionamentos

algoritmo para identificação de entidades nomeadas e o método utilizado para gerar o grafo de relacionamentos entre as entidades; na seção 4 são apresentados os resultados encontrados durante as validações feitas com dados e usuários reais; e, na seção 5, são apresentadas as considerações finais e trabalhos em andamento.

2. Agrupamento hierárquico

Para a implementação do agrupamento hierárquico foi utilizado um algoritmo da classe *Unweighted Pair Group Method with Arithmetic mean - UPGMA* [Jain et al. 1999, Manning and Schütze 2003]. Trata-se de um algoritmo *bottom-up* que usa uma função de distância euclidiana como função de similaridade. O algoritmo recebe um conjunto de objetos, iniciando com um agrupamento para cada objeto. Em cada passo, os dois agrupamentos mais similares são determinados e unidos em um novo agrupamento. Elimina-se os dois agrupamentos mais similares e adiciona-se o novo agrupamento ao conjunto de agrupamentos. O algoritmo é finalizado quando o número de agrupamentos for igual a 1. Na figura 1.a é possível visualizar a forma gráfica utilizada para representar o agrupamento hierárquico.

O pré-processamento de cada documento inclui algoritmos de *stemming*, lista de *stop-words* e a transformação de cada documento em um vetor utilizando a equação *TF-IDF* [Manning and Schütze 2003]. O algoritmo de agrupamento foi modificado para fornecer uma descrição, um rótulo, para cada agrupamento dentro da hierárquia. O algoritmo para criação dos rótulos é baseado na hipótese que uma expressão que é muito freqüente em um agrupamento e difícil de ser encontrada na coleção de documentos provavelmente será útil para a descrição do agrupamento analisado.

A seleção das expressões candidatas para a descrição do agrupamento é baseada no seguinte critério $Candidatos = \{e \mid \frac{NDc}{|C|} < max \wedge \frac{NDa}{|A|} > min\}$, onde NDc é o número de documentos na coleção que contêm a expressão e e NDa é o número de documentos no agrupamento analisado que contêm a expressão e. O número de documentos na coleção é representado por $|C|$. O número de documentos no agrupamento é representado por $|A|$. min e max são limites configuráveis entre 0 e 1.

O algoritmo para criação dos rótulos dos agrupamentos implementado neste trabalho utiliza duas abordagens para determinar o que é uma expressão. A primeira abordagem considera os conceitos de *uni-gram*, *bi-gram* e *n-gram*

[Treeratpituk and Callan 2006, Manning and Schütze 2003]. A segunda abordagem, menos usual em outros trabalhos, considera uma expressão como sendo uma entidade nomeada. As duas abordagens são avaliadas neste trabalho.

3. Algoritmo para identificação de entidades nomeadas e Grafo de Relacionamentos

A etapa de detecção de relações entre entidades faz uso de um algoritmo que reconhece entidades nomeadas em documentos não-estruturados (textos). As entidades reconhecidas pelo algoritmo são nomes de pessoas, lugares, organizações e termos relevantes para o domínio da aplicação. Todas as entidades nomeadas em todos os documentos recuperados são apresentadas ao usuário para seleção. Após a seleção dos termos que o usuário considera relevantes são gerados grafos de relacionamentos, figura 1.b. Em um tipo de grafo, cada nó é um documento e as arestas são os termos que associam um documento ao outro; em outro tipo de grafo, os nós são as entidades nomeadas e as arestas são os documentos que relacionam uma entidade nomeada a outra.

O algoritmo para identificação de entidades nomeadas implementado é baseado no proposto em [Bikel et al. 1997]. Este algoritmo faz uso de um Modelo Oculto de Markov que é treinado/criado a partir de um conjunto de documentos etiquetados. Na identificação de entidades nomeadas, os exemplos de treinamento devem ser etiquetados usando todas as entidades de interesse do ambiente de aplicação escolhido. Os termos etiquetados são nomes de pessoas (isto inclui apelidos), nomes de organizações (completo e abreviados), locais (nomes de cidades, estados, bairros, estabelecimentos comerciais, entre outros) e termos relevantes para o domínio (drogas e armas).

O uso de uma abordagem de aprendizado estatística permite a mineração de textos até em fontes de dados onde as sentenças são mal-formadas, ou seja, não seguem as regras de um determinado idioma. Exemplos são *blogs* na Internet e textos gerados a partir de escutas telefônicas.

4. Resultados

A validação do processo descrito neste artigo foi realizada utilizando dados de transcrições de escutas telefônicas, de Boletins de Ocorrência, relatórios de inquéritos policiais e notícias coletadas na Web, somando aproximadamente 300.000 documentos. O processo de validação contou com a colaboração de cinco investigadores. Durante a validação do sistema os investigadores puderam opinar livremente sobre as funcionalidades do sistema.

Foram apresentadas três versões do agrupamento hierárquico: sem rótulo; com rótulos gerados a partir de *uni-gram*, *bi-gram* e *tri-gram* encontrados nos documentos; e, com rótulos gerados a partir de entidades nomeadas encontradas nos documentos. Os usuários consideraram a versão que utiliza entidades nomeadas mais compreensível e intuitiva.

Através do uso de um algoritmo de agrupamento hierárquico e de um identificador de entidades nomeadas foi possível explicitar padrões ocultos entre os documentos e identificar relações entre entidades (pessoas, organizações e lugares). O algoritmo para identificação de entidades nomeadas implementado neste trabalho teve uma *medida F* variando entre $0,62$ e $0,84$ nos testes realizados.

5. Considerações Finais e Trabalhos em Andamento

Durante a avaliação do trabalho verificou-se que o processo que une os algoritmos de agrupamento hierárquico e de identificação de entidades nomeadas agrega valor à investigação policial. O algoritmo de agrupamento hierárquico é uma ferramenta útil para a análise exploratória dos dados e pré-seleção dos documentos que serão utilizados pelo algoritmo de identificação de entidades nomeadas. O algoritmo para identificação de entidades nomeadas é útil na geração de um grafo de relacionamentos entre entidades, que consegue sintetizar boa parte das informações envolvidas em uma investigação.

Ainda como resultado da avaliação do trabalho, percebeu-se a importância de bons rótulos para o agrupamento hierárquico. Neste trabalho foi apresentado um algoritmo que considera o resultado gerado por um reconhecedor de entidades nomeadas para a criação de rótulos em agrupamentos hierárquicos. Para o domínio da investigação policial esta abordagem mostrou-se melhor que outras abordagens tradicionais para identificação de expressões. A apresentação gráfica dos resultados permite ao investigador construir cenas visuais de investigações em andamento e, de forma interativa, explorar os dados e monitorar mudanças nos cenários da investigação.

Atualmente, o processo descrito neste trabalho está sendo validado em outros domínios de aplicação, entre eles: jornalismo investigativo, objetivando a análise do conteúdo produzido por diversas fontes jornalísticas na Internet; inteligência competitiva, visando o monitoramento da concorrência, análise das tendências de mercado e monitoramento do impacto de campanha de *marketing* através da análise do conteúdo publicado em *blogs*; pesquisa e desenvolvimento, objetivando a análise do conteúdo publicado em bases de patentes; e, na área da saúde, analisando informações sobre diagnósticos, tratamentos, medicamentos e doenças a partir de artigos científicos e conteúdo de prontuários.

Referências

Barth, F. J., Belderrain, M. C. R., Quadros, N. L. P., Ferreira, L. L., and Timoszczuk, A. P. (2007). Recuperação e mineração de informações para a área criminal. In *ENIA VI - Encontro Nacional de Inteligência Artificial. Anais do XXVII Congresso da SBC*, pages 1292–1301.

Bikel, D., Miller, S., Schwartz, R., and Weischedel, R. (1997). Nymble: a high-performance learning name-finder. In *Proceedings of ANLP-97*, pages 194–201.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

Júnior, C. M. F. and de Lima Dantas, G. F. (2006). A descoberta e a análise de vínculos na complexidade da investigação criminal moderna. Adquirido no site do Ministério da Justiça (http://www.mj.gov.br) em agosto de 2006.

Manning, C. D. and Schütze, H. (2003). *Foundations of Statistical Natural Language Processing*. MIT Press.

Treeratpituk, P. and Callan, J. (2006). Automatically labeling hierarchical clusters. In *Proceedings of the 2006 international conference on Digital government research*, pages 167–176, New York, NY, USA. ACM Press.