Uso de técnicas de recuperação e mineração de informação em atividades de Serviços de Informações do Estado

Fabrício J. Barth, Antonio Pedro Timoszczuk
Fundação Aplicações de Tecnologias Críticas - Atech. Rua do Rocio, 313, 11º andar. São Paulo – SP. Brasil.

Resumo **— Este trabalho descreve um sistema de recuperação e mineração de informação projetado para auxiliar a atividade de investigação. Este sistema é capaz de processar diversas fontes de informação e identificar relações entre os documentos recuperados. As evidências de relacionamento entre documentos são apresentadas na forma gráfica. A apresentação gráfica dos resultados permite ao investigador construir panoramas visuais de investigações em andamento e, de forma interativa, explorar os dados e monitorar mudanças nos cenários da investigação. A avaliação apresentada neste trabalho foi realizada em um ambiente de investigação criminal. No entanto, acredita-se que o sistema descrito neste trabalho seja útil também como ferramenta de apoio para a investigação realizada em Serviços de Informações do Estado.**

Palavras-chaves **— Comando e Controle, Sistemas de Informação, Recuperação e Mineração de Informações, Inteligência de Máquina.**

I. INTRODUÇÃO

Para garantir a segurança do estado, a maior parte dos governos recorrem não só à polícia mas também aos Serviços de Informações. Estes serviços têm por missão, além de outras atividades, a aquisição e tratamento de informações que permitam fazer face à qualquer perigo para a segurança do Estado [1].

As investigações realizadas pelos serviços de informações implicam lidar com elevado número de relacionamentos, fontes diversificadas e difíceis de analisar e compreender. Na maioria dos casos, as informações consideradas relevantes se mantêm ocultas, devido ao enorme volume e aparente dispersão de dados [2].

Este artigo descreve um sistema de recuperação e mineração de informação, projetado para auxiliar a atividade de investigação. Este sistema é capaz de processar diversas fontes de informação (pública, privada, estruturada e não-estruturada) e identificar relações entre documentos recuperados. As evidências de relacionamento entre documentos são obtidas através de técnicas de mineração. As técnicas de mineração utilizadas são algoritmos de agrupamento e algoritmos para identificação de entidades nomeadas. Os relacionamentos encontrados são apresentados na forma gráfica.

O objetivo deste sistema é dual: facilitar o acesso às diversas fontes de dados, através de um mecanismo de recuperação de informação que utiliza uma ontologia de domínio para gerar consultas contextualizadas, e explicitar padrões ocultos em uma grande quantidade de documentos, verificando a existência de elementos associados, identificando relações entre entidades e construindo modelos de informação sintetizada.

Este texto está estruturado da seguinte maneira: na seção 2 é apresentado o método para indexação e recuperação dos documentos considerados relevantes para a investigação; na seção 3 é apresentado o método para determinação de similaridades entre os documentos retornados; na seção 4 é descrito o método utilizado para gerar o grafo de relacionamentos entre as entidades relevantes para a investigação; na seção 5 são apresentados os resultados encontrados durante uma validação feita com dados e usuários reais em um ambiente de investigação, e; na seção 6, são apresentadas as conclusões e considerações finais.

Fabrício J. Barth, fbarth@atech.br,
Antonio Pedro Timoszczuk, antoniop@atech.br,
Tel +55-11-3040-7379, Fax +55-11-3040-7400.

II. INDEXAÇÃO E RECUPERAÇÃO DOS DOCUMENTOS

Para a aquisição de documentos (formação da base indexada) é utilizada uma forma de busca sistemática sobre: (*i*) arquivos RSS de *sites* de notícia, (*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, (*v*) conteúdo de páginas web pré-definidas por especialistas da área, e (*vi*) conteúdo de páginas web retornadas, utilizando um mecanismo de busca tradicional (por exemplo, Google e Yahoo), a partir de um conjunto de consultas formuladas por especialistas.

Para aumentar a precisão e o índice de cobertura de documentos relevantes é utilizada uma ontologia de domínio. Ao permitir a expansão da consulta do usuário a partir de termos adicionais próprios do domínio, espera-se que a ontologia viabilize a recuperação de documentos que seriam ignorados pela consulta original.

No contexto da engenharia do conhecimento, uma ontologia é especificada sob a forma de um vocabulário que representa os conceitos do domínio [3]. Um exemplo simples, que corresponde à ontologia desenvolvida neste projeto, é o da hierarquia de tipos, onde são especificadas classes (conceitos) e seus relacionamentos com superclasses e subclasses. Esse tipo de ontologia resulta, portanto, da decomposição do domínio em conceitos que se relacionam com outros mais

genéricos e mais específicos. Todos os conceitos podem ter sinônimos associados. Os sinônimos permitem a cobertura exaustiva do vocabulário do domínio, enquanto que a hierarquia de tipos permite a navegação de conceitos mais específicos para mais genéricos, e vice-versa.

A consulta submetida pelo usuário é contextualizada da seguinte maneira: cada termo que compõe a consulta é pré-processado tendo em vista a remoção de acentos, de maiúsculas e a redução ao singular. Em seguida, o termo é confrontado com aqueles definidos na ontologia e seus sinônimos. Se o termo ou um de seus sinônimos for encontrado na ontologia, todos são conectados com o operador OR, compondo uma expressão que vai substituir o termo original na consulta. Caso o termo não esteja presente na ontologia, ele é mantido sem alterações na consulta. Todos os termos, expandidos ou não, são conectados com o operador AND para formar a consulta contextualizada. Termos compostos podem fazer parte tanto da consulta como da ontologia. Tais termos, que devem ser especificados entre aspas pelo usuário, não são reduzidos ao singular.

A critério do usuário, a consulta original ou contextualizada pode ser refinada mediante duas operações: focalização e generalização [4]. Na generalização, os termos da consulta e seus sinônimos são substituídos pelos termos correspondentes à(s) superclasse(s) e seus sinônimos. Na focalização, são acrescentados aos termos da consulta e seus sinônimos os termos correspondentes à(s) subclasses(s) e seus sinônimos. O objetivo da primeira operação é obter uma consulta mais abstrata (composta de termos mais genéricos na hierarquia); o da segunda, obter uma consulta mais específica (composta de termos mais específicos na hierarquia). Isto permite ao usuário navegar a ontologia tendo em vista o ajuste de sua consulta ao nível de abstração desejado.

III. AGRUPAMENTO HIERÁRQUICO

Para o agrupamento hierárquico são utilizados algoritmos que realizam o particionamento de um conjunto de objetos. O objetivo dos algoritmos de agrupamento é colocar objetos similares em um mesmo grupo e objetos não similares em grupos diferentes. Um agrupamento hierárquico é representado por uma árvore. Os nós folhas são os objetos. Cada nó intermediário representa o agrupamento que contém todos os objetos de seus descendentes [5].

Uma distinção entre a abordagem hierárquica e as demais é que o resultado obtido não é constituído apenas de uma partição do conjunto de dados inicial, mas sim de uma hierarquia que descreve um particionamento diferente a cada nível analisado. Um conjunto de dados contém, geralmente, diversos agrupamentos, que por sua vez, contém sub-agrupamentos. Os sub-agrupamentos podem ainda ser formados a partir do agrupamento de outros agrupamentos menores, e assim sucessivamente [6].

Um aspecto positivo do agrupamento hierárquico é a flexibilidade em relação à análise dos diferentes níveis de granularidade e densidade de agrupamentos [6]. O principal uso deste algoritmo, neste domínio de aplicação, é realizar a análise exploratória dos dados recuperados.

Para a implementação do agrupamento hierárquico foi utilizado um algoritmo da classe *Unweighted Pair Group Method with Arithmetic mean – UPGMA* [7,5]. 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. Eliminam-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 é possível visualizar a forma gráfica utilizada para representar o agrupamento hierárquico.

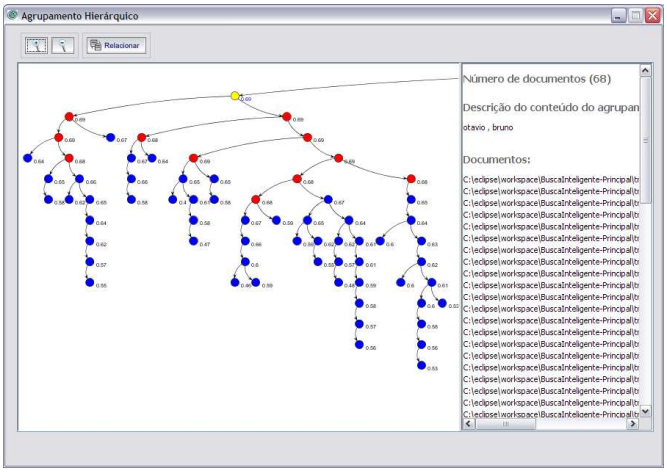

Fig. 1. Agrupamento hierárquico.

O pré-processamento de cada documento inclui algoritmos de *stemming* [8], lista de *stop-words* [5] e a transformação de cada documento em um vetor utilizando a equação *TF-IDF* (*term frequency – inverse document frequency*) [9]. O algoritmo de *stemming* consiste em uma normalização lingüística na qual as formas variantes de um termo são reduzidas a uma forma comum denominada *stem*. A conseqüência da aplicação de algoritmos de *stemming* consiste na remoção de prefixos ou sufixos de um termo, ou mesmo na transformação de um verbo na sua forma no infinitivo. Uma lista de *stop-words* é formada por um conjunto de palavras pouco significativas (conjunções, preposições e artigos) que serão removidas da descrição do documento. Usando a equação *TF-IDF* o peso do termo é proporcional ao número de ocorrências do termo no documento e inversamente proporcional ao número de documentos onde o termo aparece. Algoritmos de *stemming* e lista de *stop-words* podem ser utilizados para reduzir a dimensão dos vetores que representam os documentos.

A seleção das expressões candidatas para a descrição do agrupamento é baseada no seguinte critério definido na equação 1.

$$\text{Candidatos} = \{e \mid NDc / |C| < \max \text{ e } NDa / |A| > \min\} \quad (1)$$

Onde, *NDc* é o número de documentos na coleção que contêm a expressão *e* e *NDa* é o número de documentos no agrupamento analisados 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* [10,

5]. A segunda abordagem, menos usual em outros trabalhos, considera uma expressão como sendo uma entidade nomeada. As duas abordagens são avaliadas neste trabalho.

IV. DETERMINAÇÃO DE RELAÇÕES USANDO UM RECONHECEDOR DE ENTIDADES NOMEADAS

A etapa de detecção de relações entre documentos 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 2). 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.

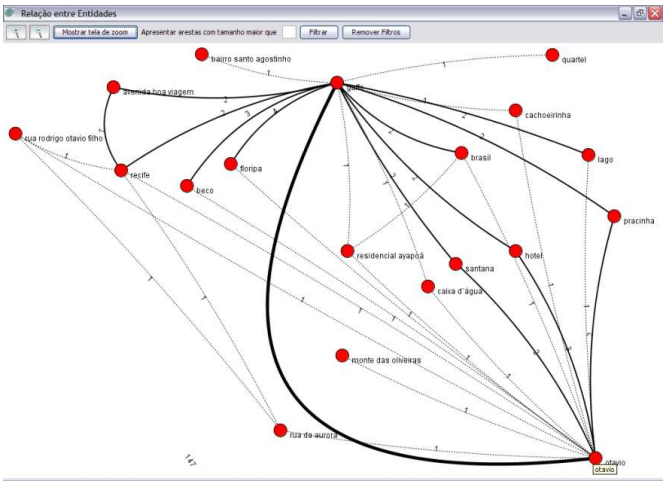

Fig. 2. Grafo de relacionamentos.

O algoritmo para identificação de entidades nomeadas implementado é baseado no proposto em [11]. Este algoritmo faz uso de um Modelo Oculto de Markov (HMM - *Hidden Markov Models*) que é treinado/criado a partir de um conjunto de documentos etiquetados. Em um documento etiquetado todas as palavras (átomos) devem ser rotuladas com uma determinada classe (i.e., pessoa, organização, lugar, entre outros). Trata-se de um processo de aprendizado supervisionado de um modelo estatístico. Pesquisas recentes demonstram a eficiência dos Modelos Ocultos de Markov nas tarefas de extração de informação [11,12]. Em muitos casos, a acurácia destes modelos é significativamente superior à de outras abordagens.

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 (completos e abreviados), locais (nomes de cidades, estados, bairros, estabelecimentos comerciais, dentre outros) e termos relevantes para o domínio (armas).

O reconhecedor de entidades nomeadas é utilizado apenas para as fontes de dados não-estruturadas. Para as fontes de dados estruturadas o acesso aos nomes, locais e organizações envolvidas é feito diretamente via banco de dados. 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. Os resultados alcançados com esta abordagem são apresentados na próxima seção.

V. RESULTADOS

A validação do sistema 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.

Durante esta validação foram avaliados todos os módulos do sistema. 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.

V.I. AVALIAÇÃO DO MÓDULO DE RECUPERAÇÃO DE DOCUMENTOS

Cada especialista envolvido na avaliação formulou cinco consultas e identificou, para cada consulta formulada, os documentos que considerava relevantes entre os 100 documentos filtrados aleatoriamente antes do início da avaliação.

Para cada consulta, foram aplicadas as três operações implementadas (contextualizar, focalizar e generalizar) e uma outra operação onde a consulta não foi alterada em função da ontologia, chamada de consulta simples. Os resultados obtidos mostram que as operações que fazem uso da ontologia (contextualizar, focalizar e generalizar) possuem, na maioria dos casos, uma medida de precisão [13] (figura 3) e de cobertura [13] (figura 4) superiores à da busca simples.

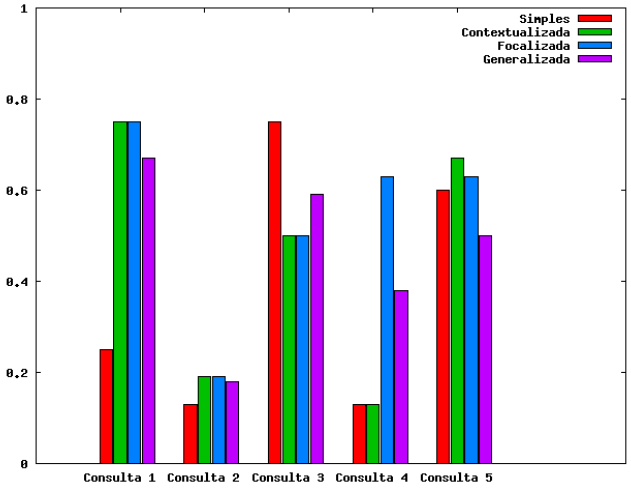

Fig. 3. Resultados da medida de precisão

A média da *medida F*, uma medida que unifica as medidas de precisão e cobertura [13], da busca simples é 0.41, enquanto que as médias da *medida F* das buscas contextualizadas, focalizadas e generalizadas são: 0.48, 0.54 e 0.52, respectivamente.

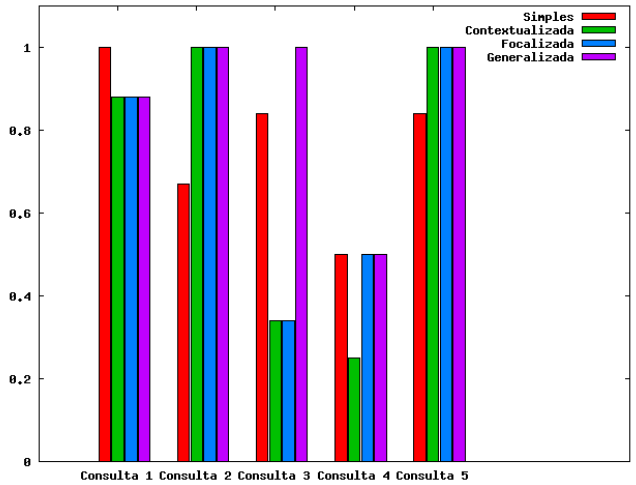

Fig. 4. Resultados da medida de cobertura

Levando-se em consideração que o objetivo da operação generalizada é obter uma consulta mais abstrata e o da operação focalizada é obter uma consulta mais específica. Em termos de medida de precisão e cobertura, isto significa dizer que: a consulta generalizada deve ter um índice de precisão mais baixo e um índice de cobertura mais alto, e; a consulta focalizada deve ter um índice de precisão mais alto e um índice de cobertura mais baixo.

Com a execução dos experimentos foi possível constatar que: a média da precisão da consulta focalizada (0.54) é maior que a média da precisão da consulta generalizada (0.46), e; a média do índice de cobertura da consulta focalizada (0.74) é menor que a média do índice de cobertura da consulta generalizada (0.88).

V.II. AVALIAÇÃO DO MÓDULO GERADOR DE AGRUPAMENTO HIERÁRQUICO

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 utilizado para cálculo do agrupamento hierárquico é um algoritmo com ordem de grandeza $O(n^2)$, onde n é o número de uniões realizadas durante o processo do agrupamento hierárquico. O número de uniões é exatamente o número de documentos utilizados menos um ($docs - 1$). Em uma máquina com processador Pentium 3 e 512 MB de memória, cada etapa do algoritmo para agrupamento hierárquico é processada em aproximadamente 1 milisegundo. Nesta situação, o tempo total de processamento de um agrupamento com 200 documentos é de 40 segundos.

V.III. AVALIAÇÃO DO MÓDULO PARA IDENTIFICAÇÃO DE ENTIDADES NOMEADAS

Na avaliação do módulo para identificação de entidades nomeadas foram utilizados três modelos distintos. O modelo m_1 foi criado/treinado apenas com o conteúdo etiquetado de notícias encontradas na Web. Os modelos m_2 e m_3 foram criados com o conteúdo etiquetado de notícias e transcrições de escutas telefônicas. As notícias e as transcrições de escutas telefônicas foram selecionadas de maneira aleatória. Na tabela 1 é apresentado o tamanho dos conjuntos de treinamento utilizados.

Modelos	*Notícias da Web*	*Escutas telefônicas*	*Total*
m_1	19.451	0	19.451
m_2	19.451	761	20.212
m_3	19.451	999	20.450

Tabela 1. Números de átomos utilizados para o treinamento dos modelos

Os testes foram realizados com três tipos de conjunto de testes: testes levando em consideração apenas notícias, testes levando em consideração notícias e escutas e testes levando em consideração apenas escutas. Na figura 5 é possível visualizar o índice de cobertura e precisão para o modelo m_1 com testes realizados apenas com notícias (*cobertura=0,85* e *precisão=0,64*). Aplicando este mesmo modelo a um conjunto de testes com notícias e escutas, os índices de cobertura e precisão são reduzidos drasticamente (*cobertura=0,32* e *precisão=0,47*). Os modelos m_2 e m_3 foram treinados com escutas e notícias: isto justifica o aumento significativo do índice de recuperação e precisão a partir do modelo m_2 (cobertura=*0,57* e *precisão=0,62*) nos testes realizados com notícias e escutas telefônicas. A diferença entre o modelo m_2 e m_3, nos testes com notícias e escutas, ocorre porque o modelo m_3 foi treinado com uma quantidade maior de escutas que o modelo m_2. Nos testes realizados apenas com notícias não existiu nenhuma variação nos índices de cobertura e precisão entre os modelos m_1, m_2 e m_3 porque não foi acrescentado nenhum conhecimento adicional sobre notícias a partir do modelo m_1.

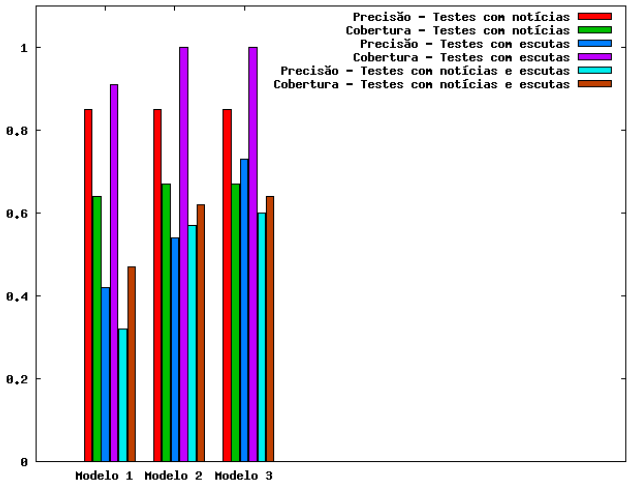

Fig. 5. Resultados da avaliação dos modelos para identificação de entidades nomeadas

VI. CONSIDERAÇÕES FINAIS

Este trabalho apresentou um sistema de recuperação e mineração de informações projetado para auxiliar a atividade de investigação. Este sistema possui dois módulos: um mecanismo de recuperação de informações que utiliza uma ontologia de domínio para gerar consultas contextualizadas,

e; um módulo para identificação de relações entre os documentos recuperados.

A análise dos resultados obtidos demonstrou que as operações que fazem uso da ontologia (contextualizar, focalizar e generalizar) possuem um desempenho superior quando comparadas a um processo de recuperação de informação convencional. Além disso, observou-se que a operação generalizada obtém uma consulta mais abstrata enquanto que a operação focalizada obtém uma consulta mais específica, permitindo ao usuário navegar a ontologia tendo em vista o ajuste de sua consulta ao nível de abstração desejado.

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 tarefa de investigação. 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. 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.

Apesar da primeira avaliação do sistema ter sido desenvolvida em um ambiente de investigação policial, acredita-se que o sistema descrito neste trabalho seja útil também como ferramenta de apoio para a investigação realizada em Serviços de Informações do Estado.

REFERÊNCIAS

[1] Comissão Temporária sobre o Sistema de Interceptação ECHELON – Parlamento Europeu. Relatório sobre a existência de um sistema global de interceptação de comunicações e econômicas. Relatório, 2001.

[2] 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.

[3] Mike Uschold and Michael Gruninger. Ontologies: Principles, methods and applications. The Knowledge Engineering Review, 11(2):93–155, 1996.

[4] Dario Bonino, Fulvio Corno, Laura Farinetti, and Alessio Bosca. Ontology driven semantic search. WSEAS Transaction on Information Science and Application, 1(6):1597–1605, December 2004.

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

[6] Metz, J. and Monard, M. C. (2006). Estudo e análise das diversas representações e estruturas de dados utilizadas nos algoritmos de clustering hierárquico. Technical report, Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo. São Carlos, São Paulo. Brasil.

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

[8] Porter, M. (1980). An algorithm for sufix stripping program. Program, 14(3): 130-137.

[9] Salton, G. and Buckley, C. (1988). Term weighting approaches in automatic text retrieval. Information Processing and Management, 24:513-523.

[10] Treeratpituk, P. and Callan, J. (2006). Automatically labeling hierarchical clusters. In Proceedings of the 2006 International Conferences on Digital Government Research, pages 167-176, New York, NY, USA. ACM Press.

[11] 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.

[12] Freitag, D. and McCallum, A. (2000). Information extraction with HMM structures learned by stochastic optimization. In AAAI/IAAI, pages 584-589.

[13] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.