Recuperação e mineração de informações para a área criminal*

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)
Rua do Rocio, 313 - 11º andar, Vila Olímpia, São Paulo - SP

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

Abstract. *This paper describes an information retrieval system prototype designed to process structured and non-structured information sources. Those sources are searched according to a query where the terms chosen by the user may be expanded when matched against a domain ontology. The retrieved documents are then submitted to clustering and named entities recognition algorithms. Both are text mining techniques which expose the relationships among the documents and allow the graphical display of those to the user. The evaluation presented here is done under a criminal information environment.*

Resumo. *Este artigo descreve o protótipo de um sistema de recuperação de informação projetado para processar fontes de informação estruturada e não-estruturada. Tais fontes são pesquisadas a partir de uma consulta onde os termos escolhidos pelo usuário podem ser expandidos mediante confronto com uma ontologia de domínio. Os documentos recuperados são então submetidos a algoritmos de agrupamento e de identificação de entidades nomeadas. Ambos são técnicas de mineração de texto que colocam em evidência as relações entre os documentos e permitem a apresentação gráfica das mesmas ao usuário. A avaliação apresentada neste trabalho foi realizada em um ambiente de investigação criminal.*

1. Introdução

As investigações policiais contemporâneas envolvem a análise de uma enorme quantidade de dados, em múltiplos formatos, originados de três fontes básicas: (i) humanas, (ii) de conteúdo, e (iii) tecnológicas. As fontes humanas podem ser determinadas nos depoimentos, interrogatórios, denúncias e entrevistas com colaboradores e informantes. As fontes de conteúdo podem ser exemplificadas com os registros provenientes de sistemas bancários, ocorrências policiais, notícias da mídia, bem como de documentos de toda ordem. Já as fontes tecnológicas têm sua expressão na telecomunicação, imagens e sinais eventualmente interceptados, captados e devidamente analisados [Júnior and de Lima Dantas 2006].

Em algum lugar, no âmago de um complexo de dados e informações provenientes de fontes humanas, de conteúdo e de tecnologia, pode estar a ”chave” de uma investigação

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

que, entretanto, se mantém oculta, devido ao enorme volume e aparente dispersão de dados e/ou informações individualmente consideradas. Assim, investigar um crime implica lidar com relações numerosas, diversificadas e difíceis de analisar e compreender. O sucesso de uma investigação criminal, portanto, irá certamente depender da capacidade de analisar e perceber, em sua complexidade, os dados distintos e seus inter-relacionamentos existentes[Júnior and de Lima Dantas 2006].

Desta forma, a investigação policial precisa ser multifacetada, dada a complexidade de seus objetos, devendo poder realizar as seguintes ações: (i) verificar a existência de elementos associados, (ii) identificar relações entre fatos conexos, e (iii) construir modelos de informação sintetizada, possibilitando a compreensão da investigação como um todo e de suas partes constitutivas. Assim, 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], constituindo um ambiente ideal para a aplicação de ferramentas avançadas de recuperação e mineração de dados.

Este trabalho descreve um protótipo de um sistema de recuperação de informação projetado para processar fontes de informação estruturada e não-estruturada, sejam elas públicas ou privadas. Tais fontes são pesquisadas a partir de uma consulta onde os termos escolhidos pelo usuário podem ser expandidos mediante confronto com uma ontologia de domínio. Os documentos recuperados são então submetidos 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 os documentos e permitem a apresentação gráfica das mesmas ao usuário (figuras 1 e 2, seção 5).

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 reconhecimento de entidades nomeadas e como estas entidades são utilizadas para gerar o gráfico de relacionamentos entre os documentos; na seção 5 são apresentados os resultados encontrados durante uma validação feita com dados e usuários reais em um ambiente de investigação criminal; e, na seção 6, são apresentadas as conclusões e considerações finais.

2. 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 recuperação 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, a ontologia viabiliza 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. Um exemplo simples, que corresponde à ontologia desenvolvida inicialmente neste projeto, é o da hierarquia de tipos, onde são especificadas classes (conceitos) e seus relacionamentos com superclasses e subclasses [Uschold and Gruninger 1996].

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 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 [Bonino et al. 2004]. 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 concreta (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.

A partir de um conjunto de documentos recuperados, são utilizadas duas abordagens para identificação de relações entre os documentos: agrupamento hierárquico sobre os documentos e relações entre documentos utilizando entidades nomeadas.

3. Agrupamento hierárquico

Para o agrupamento hierárquico são utilizados algoritmos que realizam o particionamento de um conjunto de objetos [Manning and Schütze 2003]. 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 [Manning and Schütze 2003]. Os nós folhas são os objetos. Cada nó intermediário representa o agrupamento que contém todos os objetos de seus descendentes.

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 [Metz and Monard 2006].

Um aspecto positivo do agrupamento hierárquico é a flexibilidade em relação à análise dos diferentes níveis de granularidade e densidade de agrupamentos [Metz and Monard 2006]. 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* [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 2, na seção 5, é possível visualizar a forma gráfica utilizada para representar o agrupamento hierárquico.

O pré-processamento de cada documento inclui algoritmos de *stemming* [Porter 1980], lista de *stop-words* [Manning and Schütze 2003] e a transformação de cada documento em um vetor utilizando a equação *TF-IDF* [Salton and Buckley 1988]. 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.

4. 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 reconheci-

das 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, seção 5). 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. Em um documento etiquetado todas as palavras (átomos) devem ser rótuladas 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 (HMM[1]) nas tarefas de extração de informação [Freitag and McCallum 2000]. Os Modelos Ocultos de Markov têm sido aplicados com sucesso na tarefa de reconhecimento de entidades nomeadas [Bikel et al. 1997]. 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 (completo e abreviados), locais (nomes de cidades, estados, bairros, estabelecimentos comerciais, entre outros) e termos relevantes para o domínio (drogas e 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.

5. Resultados

A validação do sistema descrito neste artigo foi realizada utilizando dados de transcrições de escutas telefônicas, de Boletins de Ocorrência e notícias coletadas na Web. O processo de validação do sistema contou com a colaboração de cinco investigadores e dois especialistas da área criminal.

A validação foi dividida em duas etapas: uma etapa chamada de qualitativa, com a participação de investigadores, e uma outra chamada de quantitativa, com a participação de especialistas da área criminal. Na etapa qualitativa, os objetivos foram: (i) verificar se o sistema agrega valor ao processo de investigação, e (ii) identificar as mudanças a serem feitas no sistema para uma melhor adequação ao processo de investigação.

Os objetivos da avaliação quantitativa foram: (i) avaliar o índice de precisão[2] e

[1]do inglês, *Hidden Markov Models*

[2](Número de documentos relevantes e recuperados) / (Número de documentos recuperados)

recuperação[3] do algoritmo de recuperação de informação descrito na seção 2, (ii) verificar a qualidade do modelo e do algoritmo para identificação de entidades nomeadas descrito na seção 4, e (iii) mensurar o tempo para cálculo do agrupamento hierárquico, descrito na seção 3.

5.1. Avaliação qualitativa

Durante a validação, as funcionalidades do sistema foram demonstradas com a utilização dos dados locais, ou seja, transcrições de escutas ou boletins de ocorrência com conteúdo conhecido para os investigadores - somando aproximadamente 200.000 documentos. Durante a demonstração das funcionalidades os investigadores puderam opinar livremente sobre o sistema.

Inúmeros sinônimos foram acrescentados à ontologia. Por exemplo, *"whisky"* é um termo muito utilizado nas ligações telefônicas e é sinônimo de entorpecente. *"Pedra"*, *"branca"* e *"feijão"* também são termos muito utilizados em ligações telefônicas e são sinônimos para cocaína e maconha, respectivamente. A ontologia utilizada na avaliação possui aproximadamente 80 conceitos. Cada conceito com 3 sinônimos em média.

Na figura 1 é possível visualizar um grafo de relacionamentos. Nesta figura existem vários nomes de pessoas: $pessoa_1, pessoa_2, \cdots, pessoa_{12}$; nomes de lugares: *"igreja"*, *"praça"* e *"bar do miguel"*; e entidades relevantes para o domínio: *"pedra"* e *"whisky"*[4].

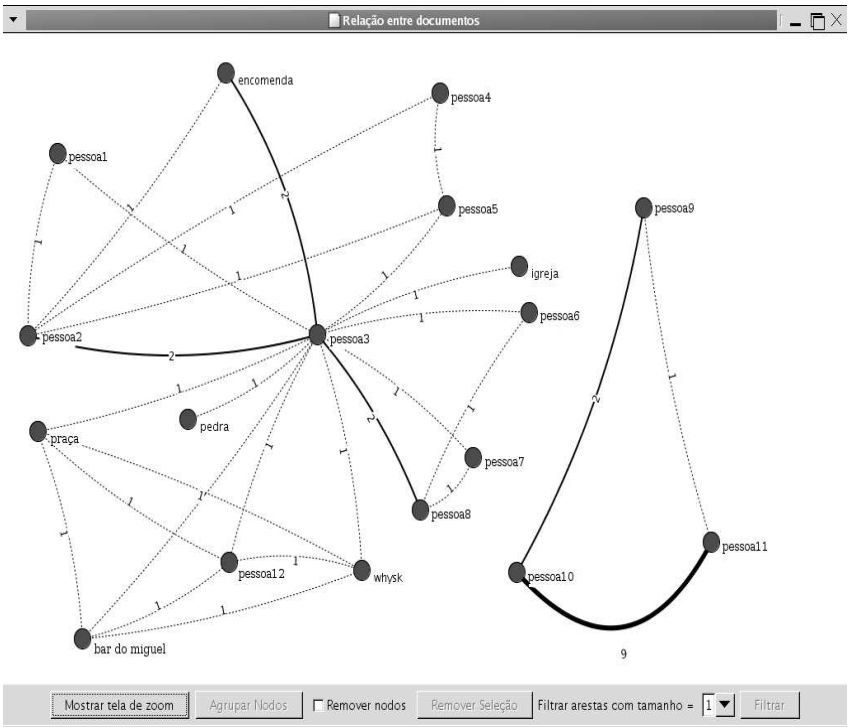

Figura 1. Exemplo de grafo de relacionamentos

[3](Número de documentos relevantes e recuperados) / (Número de documentos relevantes)

[4]Para não divulgar dados sigilosos, os verdadeiros nomes de pessoas, lugares e organizações foram trocados por nomes fictícios. Demais informações foram mantidas.

Percebe-se no grafo (figura 1) que existe uma forte ligação entre $pessoa_{11}$ e $pessoa_{10}$ e que $pessoa_9$ é uma pessoa ligada à $pessoa_{11}$ e à $pessoa_{10}$. Ao comentar estas percepções, os investigadores imediatamente notaram que a $pessoa_{10}$ é mulher da $pessoa_{11}$ e que a $pessoa_9$ é advogada da $pessoa_{10}$. A $pessoa_{11}$ está presa e a $pessoa_9$ está resolvendo os problemas relacionados ao carro da $pessoa_{10}$.

Ao aplicar o algoritmo de agrupamento a um conjunto de documentos, foi possível verificar que algumas escutas de números de telefones distintos têm um alto grau de similaridade. Isto pode ser visualizado na figura 2. No canto direito da tela são visualizadas duas escutas de telefones diferentes, usados pelo mesmo traficante, que fala com a mesma pessoa ($pessoa_{10}$). Em uma das escutas, o investigador não identifica o nome do interlocutor, mas é possível inferir que é o $pessoa_{11}$ porque este está falando com a $pessoa_{10}$.

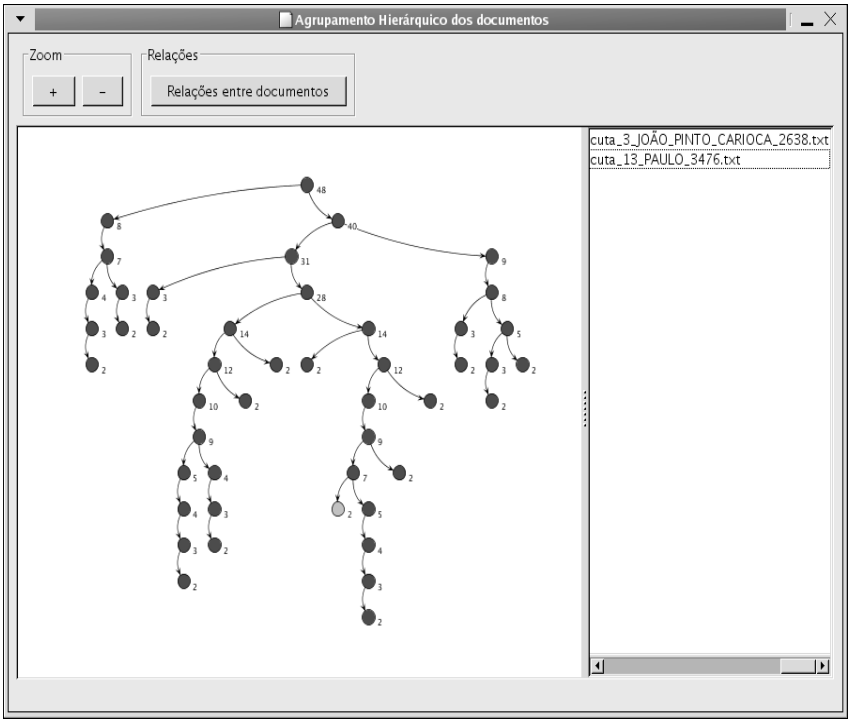

Figura 2. Exemplo de agrupamento hierárquico

Percebeu-se a necessidade de realizar consultas disjuntivas, compostas apenas pelo operador OR. Por exemplo, o investigador quer descobrir se existem relacionamentos entre o $pessoa_3$ e o $pessoa_{11}$. Ou seja, o investigador quer selecionar todos os documentos onde apareça $pessoa_3$ ou $pessoa_{11}$.

5.2. Avaliação quantitativa

A avaliação do algoritmo de recuperação de informação contou com a colaboração de dois especialistas da área criminal. Cada especialista envolvido na avaliação formulou cinco consultas e identificou, para cada consulta formulada, os documentos que considerava relevantes entre os 100 documentos filtrados aleatóriamente 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 valores dos índices de precisão,

recuperação e *medida F*[5] para cada operação (simples, contextualizada, focalizada e generalizada) são apresentados na tabela 1.

Tabela 1. Número de documentos considerados relevantes pelo usuário (D), índices de precisão (P), recuperação (R) e medida F (F) para o mecanismo de recuperação de documentos

Consultas	Usuário 1				Usuário 2			
	D	P	R	F	D	P	R	F
Simples (1)	1	0,2	1	0,33	4	0,29	1	0,45
Contextualizada (1)	1	0,5	1	**0,67**	4	1	0,75	**0,86**
Focalizada (1)	1	0,5	1	**0,67**	4	1	0,75	**0,86**
Generalizada (1)	1	0,33	1	0,5	4	1	0,75	**0,86**
Simples (2)	1	0,03	1	0,06	3	0,22	0,33	0,26
Contextualizada (2)	1	0,13	1	**0,23**	3	0,25	1	**0,4**
Focalizada (2)	1	0,13	1	**0,23**	3	0,25	1	**0,4**
Generalizada (2)	1	0,11	1	0,2	3	0,25	1	**0,4**
Simples (3)	1	0,5	1	0,67	3	1	0,67	**0,8**
Contextualizada (3)	1	0	0	0	3	1	0,67	**0,8**
Focalizada (3)	1	0	0	0	3	1	0,67	**0,8**
Generalizada (3)	1	1	1	**1**	3	0,17	1	0,29
Simples (4)	2	0,25	1	**0,4**	2	0	0	0
Contextualizada (4)	2	0,25	0,5	0,33	2	0	0	0
Focalizada (4)	2	0,25	0,5	0,33	2	1	0,5	**0,67**
Generalizada (4)	2	0,25	0,5	0,33	2	0,5	0,5	0,5
Simples (5)	1	0,2	1	0,33	6	1	0,67	0,8
Contextualizada (5)	1	0,33	1	**0,5**	6	1	1	**1**
Focalizada (5)	1	0,25	1	0,4	6	1	1	**1**
Generalizada (5)	1	0,13	1	0,23	6	0,86	1	0,92

Os resultados obtidos mostram que as operações que fazem uso da ontologia (contextualizar, focalizar e generalizar) possuem, na maioria dos casos, uma *medida F* superior à da busca simples.

Na avaliação do algoritmo 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 2 é apresentado o tamanho dos conjuntos de treinamento utilizados.

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 tabela 3 é possível visualizar o índice de recuperação e precisão para o modelo m_1 com testes realizados apenas com notícias

[5]Medida-F = (2 * precisão * recuperação) / (precisão + recuperação)

Tabela 2. Números de átomos utilizados para o treinamento dos modelos

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 3. Resultados da avaliação do modelo para identificação de entidades nomeadas (P = precisão, R = recuperação, F = medida F)

Modelos	Testes com notícias			Testes com escutas			Testes com notícias e escutas		
	P	R	F	P	R	F	P	R	F
m_1	0,85	0,64	0,73	0,42	0,91	0,58	0,32	0,47	0,38
m_2	0,85	0,67	0,75	0,54	1,00	0,70	0,57	0,62	0,60
m_3	0,85	0,67	**0,75**	0,73	1,00	**0,84**	0,60	0,64	**0,62**

(*recuperação=0,85 e precisão=0,64*). Aplicando este mesmo modelo a um conjunto de testes com notícias e escutas, os índices de recuperação e precisão são reduzidos drasticamente (*recuperação=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 (*recuperação=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 recuperação 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.

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.

6. Conclusões e Considerações Finais

Este trabalho apresentou um mecanismo de recuperação de informações que utiliza uma ontologia de domínio para gerar consultas contextualizadas aplicado a um ambiente de investigação criminal. 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.

Através do uso de um algoritmo de agrupamento hierárquico e de um identificador de entidades nomeadas foi possível explicitar padrões ocultos entre os documentos recuperados e identificar relações entre entidades (pessoas, organizações e lugares). Durante a avaliação qualitativa 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.

O algoritmo para identificação de entidades nomeadas implementado neste trabalho teve uma *medida F* variando entre $0,62$ e $0,84$ nos testes realizados. Levando-se em consideração o pequeno conjunto de treinamento utilizado, apenas 20.450 átomos, o resultado alcançado foi muito bom. Para alcançar um desempenho superior seria necessário etiquetar um conjunto de treinamento maior. Um dos trabalhos em andamento é a criação/treinamento de um modelo para identificação de entidades nomeadas através de uma abordagem semi-supervisionada, tendo em vista a utilização de uma abordagem menos laboriosa para a criação do modelo.

Agradecimentos

Agradecemos ao Dr. Francisco Sá Cavalcante, titular da Secretaria de Segurança Pública do Estado do Amazonas, e Thomaz Augusto de V. Dias, diretor do Departamento de Inteligência, pela participação neste projeto. Agradecemos também a Adriana Simizo, Bruno Furtado, Carlos Fidalgo e Mário Corrêa pelas contribuições realizadas ao trabalho.

Referências

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.

Bonino, D., Corno, F., Farinetti, L., and Bosca, A. (2004). Ontology driven semantic search. *WSEAS Transaction on Information Science and Application*, 1(6):1597–1605.

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

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.

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.

Porter, M. (1980). An algorithm for suffix stripping program. *Program*, 14(3):130–137.

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

Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*.