Mineração de Regras de Associação em Textos: Uma Aplicação em Segurança Pública

Luiz Filho[1],Eloi Favero[1] and Maxwel Dias[1]

[1] Universidade Federal do Pará, Programa de Pós Graduação em Ciência da Computação.
Belém-PA, Brazil
{ lasf.bel,maxwelmdias}@gmail.com, favero@ufpa.br

Abstract. Nowadays, it is observed an increase in the amount of unstructured data in the databases of the organizations. More and more, it is needed techniques that get to transform those data in important information for the decision taking. The Text Mining has been providing satisfactory results in the knowledge discovery in unstructured data. This paper presents an application of Knowledge Discovery in Textual Databases (KDT) techniques of security public data Para's State using Association Rules. Among the found results, it is observed that for the rape crime, firearms are used by the victim's unknown people to force them going in a place for make the crime, different from the crimes practiced by the victims' well-known people that generally use the physical force to force them practice it the sexual relationship.

Keywords: Knowledge Discovery in Texts, Text Mining, Association Rules, Public Security.

1 Introdução

As últimas décadas apresentaram um aumento considerável na quantidade de dados que são armazenadas em formato eletrônico. O valor destes dados armazenados está tipicamente ligado à capacidade de extrair informações de mais alto nível que se encontra subjacente a estes dados, ou seja, informação útil para dar suporte à tomada de decisão.

O Processo de Descoberta de Conhecimento em Base de Dados (do inglês, *Knowledge Discovery in Database - KDD*) é definido como o processo de descoberta de novos conhecimentos, sejam padrões, tendências, relações, associações, probabilidades ou fatos, que não são óbvios ou de fácil identificação em base de dados.

Grande parte das informações das instituições, entretanto, tem sido encontrada na forma textual e descrita em linguagem natural. Assim, pesquisas recentes mostram que, pela naturalidade do formato, cerca de 80% do conteúdo on-line está em formato textual [3]. Esse tipo de informação, que pode ser facilmente encontrada na internet ou nas próprias empresas, na forma de documentos eletrônicos, não é diretamente utilizado pelas ferramentas de Mineração de dados tradicionais devido ao fato de não possuir uma estrutura como a encontrada nos bancos de dados relacionais, o que minimiza seu potencial.

Dessa forma, surgiu, recentemente, uma variante do KDD, chamado Descoberta de Conhecimento em Textos (do inglês, *Knowledge Discovery in Textual Databases - KDT*) definida como o processo de extrair padrões ou conhecimentos, interessantes e não-triviais, a partir de um conjunto de documentos textuais. O KDT destina-se ao descobrimento de padrões em textos de linguagem natural que possam revelar conhecimento útil, ou seja, aplicável à tomada de decisão [16].

Na área de segurança pública, especificamente, referente ao Estado do Pará, a análise de informações sobre a criminalidade é realizada exclusivamente a partir dos dados registrados no sistema de ocorrências policiais. No entanto, nesse mesmo sistema, existe um campo textual, onde o funcionário da delegacia descreve, em detalhes, como ocorreu o crime. Neste campo textual, muitas informações importantes são registradas que não constam nos campos estruturados, apesar disso, devido a dificuldade em ler manualmente os textos de cada ocorrência, a análise das informações textuais não são realizadas.

A partir da base de dados desse sistema de segurança pública, foram publicados em [14] e [15] a aplicação do processo de KDD utilizando a técnica de Mineração de Dados conhecida com Árvore de Decisão. Este artigo mostra uma evolução dos trabalhos anteriores desses autores na área de descoberta de conhecimento e tem como objetivo mostrar a aplicação do processo de KDT realizando a extração de Regras de Associação a partir dos textos da área de segurança pública.

Além desta Seção introdutória, este artigo está organizado da seguinte maneira: na Seção 2 é detalhada as etapas do KDD; na Seção 3 é apresentada a tarefa de Associação; na Seção 4 é mostrado o Estudo de Caso a partir da base textual da segurança pública; na Seção 5, são mostrados alguns trabalhos relacionados e, finalmente, na Seção 6 são realizadas as considerações finais e propostas para trabalhos futuros.

2 Processo de KDT

No trabalho de [4] foi utilizado inicialmente o termo KDT para indicar o processo de descoberta de conhecimento em dados não-estruturados, enquanto o termo Mineração de Textos é usado para definir uma das etapas do processo de KDT relacionada com a extração de padrões de dados textuais.

As principais áreas de conhecimento que compõem a Mineração de Textos são: Aprendizado de Máquina, Processamento de Linguagem Natural, Estatística, Inteligência Computacional, Recuperação de Informação e Mineração na Web (do inglês, *Web Mining*).

Na Figura 1 são analisadas e discutidas as etapas do KDT [1]. A sequência como mostrado na Figura 1 é uma tendência encontrada nos recentes trabalhos como, por exemplo, [11], [8] e [2].

Fig. 1. Etapas do processo de KDT [1].

Na Coleta é realizado o processo de busca e recuperação de textos, tendo como finalidade formar a base textual da qual se pretende extrair algum tipo de conhecimento.

Pré-processamento tem como objetivo prover alguma formatação e representação da massa textual. Consiste nas fases de tokenização (análise léxica), eliminação de termos considerados irrelevantes (*stopwords*), bem como a normalização morfológica dos termos (*stemming*).Tokenização tem como finalidade extrair unidades mínimas de texto a partir de um texto livre. Cada unidade é chamada de *token* e que, na grande maioria das vezes, corresponde a uma palavra do texto, podendo também estar relacionado a mais de uma palavra, símbolo ou caractere de pontuação. As palavras conhecidas como *stopwords* são aquelas que não fazem diferença quando indexadas, somente aumentando o tamanho do arquivo de índice. Exemplos de *stopwords* são artigos, preposições, conjunções, e até mesmo alguns verbos. Uma palavra em um texto pode assumir formas variadas, como por exemplo: plural, formas de gerúndio e sufixos temporais. O processo de s*temming* consiste na remoção dessas variações, sendo o resultado obtido chamado de *stem* (raiz).

A Indexação é o processo que organiza todos os termos adquiridos a partir de fontes de dados, facilitando o seu acesso e recuperação. Uma boa estrutura de índices garante rapidez e agilidade ao processo, tal como funciona o índice de um livro. O Modelo Booleano é uma das representações de documentos mais clássicas utilizadas em Mineração de Textos. Essa abordagem avalia a presença ou ausência do termo no documento, sendo binário, isto é, {0,1} os pesos atribuídos a esses termos.

Segundo [16], Mineração de Textos é definida como o processo de extrair padrões ou conhecimentos, interessantes e não-triviais, a partir de um conjunto de documentos textuais.

A etapa de Análise deve ser executada por indivíduos que, normalmente, estão interessadas no conhecimento extraído e que devem tomar algum tipo de decisão apoiada no processo de Mineração de Texto.

3 Regras de Associação

Para a aplicação da Mineração de Textos, pode-se aplicar as tarefas de Classificação, Agrupamento ou Associação que são comuns tanto na Mineração de Textos quanto na Mineração de Dados ou aplicar tarefas específicas da primeira, como a Sumarização e Extração de Características. Neste artigo, é descrita a tarefa de Associação que foi a tarefa escolhida para ser utilizada na Mineração de Textos.

A tarefa de Associação, geralmente representada por Regras de Associação, caracteriza o quanto a presença de um conjunto de itens de registros de uma base de dados implica na presença de algum outro conjunto distinto de itens dos mesmos registros [6]. Desse modo, o objetivo das Regras de Associação é encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados.

O algoritmo mais utilizado para a descoberta de Regras de Associação é o *Apriori*, tendo como principal objetivo fazer uma busca completa na base de dados de maneira que sejam obtidas as regras com os melhores valores de suporte e confiança, que são dois importantes parâmetros utilizados em Regras de Associação. Por exemplo, sendo A e B dois itens na base de dados, suporte e confiança são definidos como:

(1) **Suporte:** o suporte para um item A é proporcional ao número de registros no banco que contenham A. Isto é:

$$\text{Suporte}(\{A\}) = \frac{\text{número de registros que contém A}}{\text{total de registros na base de dados}}$$

(2) **Confiança:** a confiança (A=B) é a medida de precisão da regra, a qual é determinada pela porcentagem de registros na base de dados que contém A e também contem B. Isto é:

$$\text{Confiança}(A=B) = \frac{\text{número de registros contendo A e B}}{\text{total de registros contendo A}}$$

Na Mineração de Textos, especificamente, esta tarefa consiste na descoberta de relacionamentos existentes entre termos presentes nos documentos. De forma a identificar os termos que estão mais relacionados nos documentos analisados. Como resultado, tem-se regras que mostram quanto a presença de um termo no conjunto de documentos implica na presença de algum outro termo distinto no mesmo conjunto de documentos analisados, a partir dos valores de confiança e suporte definidos pelo usuário.

4 Estudo de Caso

Geralmente, a geração de estatísticas para apoiar os gestores de segurança pública do Estado do Pará é retirada a partir das informações cadastradas no sistema onde são realizados os registros das ocorrências policiais ocorridas em todo Estado. Cada ocorrência cadastrada possui diversos campos, tais como: local onde ocorreu o delito, meio empregado pelo infrator, dia e hora da ocorrência, descrição do crime, dentre outros. Além desses, possui ainda, o relato da ocorrência, um campo textual, onde o funcionário da delegacia descreve em detalhes como ocorreu o crime.

Neste sistema, em cada ocorrência cadastrada, muitas informações textuais são descritas no relato da ocorrência, informando os detalhes de como ocorreu o crime. Além disso, uma grande quantidade de informações que possuem campos específicos

para serem preenchidas é negligenciada e são descritas no relato da ocorrência pelo funcionário da delegacia, ou seja, determinadas informações, estão presentes apenas no relato da ocorrência. Apesar disso, as análises estatísticas e comportamentais do fenômeno da criminalidade são analisadas apenas com os relacionamentos entre as variáveis (dados estruturados), desconsiderando as informações textuais.

Dentre os diversos tipos de delitos existentes na base de dados, optou-se por trabalhar com ocorrências do crime de estupro. Esta escolha foi realizada pelo fato de que a descrição deste tipo de crime ser mais detalhada no relato da ocorrência do que outros. Além disso, existe a motivação social, uma vez que a incidência desse tipo de crime tem crescido nos últimos anos, necessitando, dessa forma, procurar entender mais a prática desse crime a fim de poder utilizar o conhecimento obtido para prevenir que novos casos ocorram.

Nesta Seção, são descritas detalhadamente cada etapa do processo de KDT a partir da base textual da segurança pública.

4.1. Coleta

Na etapa da Coleta, foram selecionadas as ocorrências do crime de estupro ocorridas nos anos de 2006 e 2007, totalizando 820 registros, porém nem todos os textos foram aproveitados, pois alguns não possuíam informação armazenada. Dessa forma, foram analisados 800 documentos.

4.2. Pré-processamento

Na etapa de Pré-processamento são utilizadas algumas das classes do Lucene [7]. O Lucene é uma API que contém diversas classes implementas em Java que executam atividades de Mineração de Textos.

Uma das classes utilizada, foi a classe *BrazilianAnalyzer* que é uma classe específica para realizar o pré-processamento de textos em português (Brasil), a qual para cada arquivo de entrada processado, realiza as seguintes atividades, seqüencialmente: Geração dos *tokens* (termos) que ocorrem em cada documento; Eliminação dos sinais de pontuação e caracteres especiais. Esta classe não elimina os números presentes nos documentos, porém foi criado um método dentro desta classe a fim incluir esta opção; Conversão de todas as palavras para minúscula; Remoção das *stopwords*. Esta classe possui uma lista *default* de *stopwords*, no entanto, verificou-se que esta lista estava incompleta, nesse sentido, pesquisou-se uma relação de classes morfológicas da língua portuguesa e construiu-se uma nova tabela de *stopwords*. Esta nova tabela possui 382 termos.

Uma vez eliminadas as *stopwords*, foi aplicado o processo de *stemming*. O Lucene possui duas classes para realizar esse processo: *BrazilianStemFilter* e *BrazilianStemmer*. Ambas as classes são aplicáveis somente a textos da língua portuguesa. Na Figura 2, pode-se observar um exemplo de arquivo textual utilizado e na Figura 3, pode-se observar o referido arquivo após a etapa de Pré-processamento.

A referida relatora, de 19 anos de idade, em dia e hora supra citados, encontrava-se dormindo em sua residência, quando foi subitamente acordada por um homem que colocava a mão na boca dela, o mesmo estava armado com um revólver. Ressalta que o referido homem a estuprou violentamente, sendo que não consegue fazer o retrato falado do referido homem, pois o mesmo apagou todas as luzes da casa. A relatora ao sentir que já estava sozinha em seu quarto saiu em desespero ao quarto da genitora a qual já vinha ao seu encontro. O assaltante adentrou na residência por uma pequena janela de vidro que fica na cozinha da casa. O assaltante levou da residência os seguintes objetos: Um aparelho celular V3 Motorola, várias peças de roupas e um aparelho de DVD. Registra para providências.

Fig. 2. Exemplo de Arquivo Textual do Crime de Estupro.

refer rela ano idad dia hor supr cit encontr dorm resid subit acord hom coloc mao boc del est arm revolv ressalt refer hom estupr violent send nao consegu retrat fal refer hom apag luz cas rela sent ja est so saiu desesper geni ja vinh encontr assalt adentr resid pequen janel vidr fic co cas assalt lev resid seguint objet aparelh celul v3 motorol var pec roup aparelh dvd registr provid

Fig. 3. Exemplo de Texto Após a Etapa de Pré-processamento.

4.3. Indexação

Para realizar a indexação dos documentos, utilizaram-se algumas classes específicas da API Lucene, a saber: *Document*, *Directory*, *Analyzer* e *IndexWriter*. Estas classes foram úteis em identificar as palavras existentes nos textos da base de dados e adicioná-las em um índice. Pode-se utilizar essas classes para analisar a quantidade de termos e a freqüência deles tendo como referência todos os documentos ou analisar a quantidade de termos e a freqüência de termos por documento.

Uma vez realizado a indexação de todos os documentos, foi necessário preparar a base de dados (nesse momento, tem-se dados estruturados, por isso denominou-se dados) conforme o padrão de arquivo de entrada para a ferramenta Weka (ferramenta que é utilizada na Mineração de Textos) que possui diversos algoritmos para a realização da Mineração de Textos. Para isto, foi necessário construção e implementação de um algoritmo a fim de gerar o arquivo ARFF (arquivo de entrada do Weka) da estrutura de indexação utilizada (Modelo Booleano).

Esse algoritmo foi implementado em Java, e o seu funcionamento consiste em considerar cada documento como uma transação. Inicialmente gerou-se uma tabela onde as linhas representam os documentos e as colunas representam os termos presentes nos documentos. Para cada célula da tabela é atribuído o valor “1” se o termo está presente no documento, senão é colocado o valor “?” (esse símbolo é utilizado pelo Weka para representar informação ausente). Ressalta-se que foi utilizado outro valor no lugar do “?” (“0”, por exemplo), pois na execução do algoritmo de Regras de Associação, o “0” seria considerada uma instância de um atributo.

Antes de gerar o arquivo ARFF a partir da tabela, o algoritmo solicita um valor mínimo de freqüência que os termos devem possuir. A definição desse valor é

importante para reduzir a quantidade de termos que serão utilizados pelo algoritmo definido pelo usuário, dessa forma, trabalha-se com os termos que ocorrem mais freqüência nos documentos.

A Figura 4 mostra um exemplo de arquivo ARFF gerado a partir do algoritmo implementado em java. Para simplificar, neste exemplo, foi utilizado uma amostra da base de dados composta por 5 documentos e o valor de freqüência mínima igual a 3. Pode-se verificar que cada atributo possui apenas um valor (indicado entre "{" e "}"), a saber o valor "1". Neste exemplo, pode-se observar que apenas três termos (acim, dia e relat) possuem no mínimo três ocorrências.

```
@relation  textMining

@attribute acim {1}
@attribute dia {1}
@attribute rela {1}

@data
1,1,1
1,?,?
1,1,1
1,1,?
?,1,1
```

Fig. 4. Exemplo no Weka de Arquivo ARFF Gerado.

4.4. Mineração de Textos

Na ferramenta Weka, foi aplicado o algoritmo *Apriori*, bem como utilizado o valor 0,01 para o suporte e o valor 0,75 para a confiança e escolheu-se o valor "8" para o número mínimo de freqüência dos termos. Com esses parâmetros, foram selecionados 1.007 termos, porém, muitos desses termos eram verbos ou substantivos poucos representativos para este domínio. Nesse sentido, selecionaram-se, no próprio Weka, os termos considerados importantes a partir da análise do especialista, resultando em um total de 221 termos.

Para mostrar como é realizada a leitura de uma Regra de Associação gerada pela ferramenta Weka, analisar-se-á a regra: "doming=1 13 ==> cas=1 11 conf:(0.85)", a partir dela, observa-se que dentre as 13 ocorrências onde o termo "doming" aparece, em 11 vezes também aparece o termo "cas", ou seja 85% (0,85). A confiança é calculada dividindo 11/13 que é igual a 0,85. Enquanto que o suporte é calculado dividindo 11/800 que é igual a 0,013, onde 800 são o total de documentos utilizados neste estudo de caso.

4.5. Análise

Nessa subseção são discutidos os principais resultados encontrados com a aplicação da Mineração de Textos. Para exemplificar, foi é realizada a análise de três regras geradas:

Regra 1: amig=1 famil=1 11 ==> cas=1 11 conf:(1). Esta regra mostra que 100% dos casos onde apareceu o termo “amigo” e “família” também apareceu o termo “casa” (suporte igual a 0,013). Esta regra revela uma importante característica do perfil do agressor do crime de estupro, que muitas vezes é conhecido da família e pratica o delito na própria casa da vítima. A partir das informações dessa regra, pode-se tomar medidas informativas a fim de conscientizar a sociedade sobre esta situação em particular. Muito já foi noticiado em jornais ou revistas mostrado o envolvimento entre conhecidos e a vítima no caso desse crime, porém, no caso dos resultados deste trabalho, tem-se informações comprovadas e relevantes a partir de uma base de dados real e, dessa forma, podem ser utilizadas para respaldar a emprego dos recursos financeiros e humanos do Estado para conscientizar a população a denunciar esse tipo de crime, mesmo envolvendo pessoas conhecidas.

Regra 2: cas=1 escol=1 estud=1 9 ==> filh=1 9 conf:(1). Esta regra mostra que 100% dos casos onde apareceu o termo “casa”, “escola” e “estudar” também apareceu o termo “filha” (suporte igual a 0,011). Esta regra mostra que o delito estupro geralmente é praticado também no trajeto da criança ou adolescente da escola para sua casa. O termo “filha” apareceu muitas vezes, pois geralmente quando a vítima possui menos de 18 anos, a mãe dela é quem a leva na delegacia e quem narra o fato. A partir das informações desta regra, pode-se utilizar medidas para conscientizar crianças e adolescentes a não irem sozinhos a lugares com pessoas conhecidas ou desconhecidas quando saírem da escola. Além disso, pode-se aumentar o efetivo policial próximo às escolas, visando inibir possíveis ações criminosas.

Regra 3: estupr=1 abandon=1 15 ==> cas=1 14 conf:(0.93). Esta regra mostra que 93% dos casos onde apareceu o termo “estupro” e “abandonada” também apareceu o termo “casa” (suporte igual a 0,017). Esta regra mostra que o acusado de cometer delito estupro, muita das vezes, procura casas abandonadas para a prática de seus delitos. A partir das informações dessa regra, pode-se fiscalizar as casas que encontram-se abandonadas para verificar se elas não estão sendo utilizadas por criminosos para cometerem seus crimes. Além disso, pode-se conscientizar a população a observar atitudes suspeitas em casas ou locais que se encontram abandonadas a fim de realizarem a denúncia para as autoridades competentes.

5 Trabalhos Relacionados

Considerando o elevado crescimento na quantidade de dados não-estruturados nos últimos anos nas bases de dados das organizações, diversas pesquisas foram e estão sendo desenvolvidas no sentido de obter informações a partir de textos.

No trabalho de [9] é abordado o desafio do Processamento de Línguas Naturais e, em especial, da Língua Portuguesa, no âmbito da Ciência da Computação e suas disciplinas, de forma a permitir o acesso participativo e universal do cidadão brasileiro ao conhecimento, da gestão da informação em grandes volumes de dados multimídia distribuídos e dos problemas complexos e interdisciplinares da modelagem computacional de sistemas artificiais, naturais e sócio-culturais. [13] mostram um estudo da aplicação de técnicas de Mineração de Dados a partir de textos da internet. Os autores abordam o problema da classificação de textos, que no caso das aplicações oriundas da internet, se torna um desafio pela própria ambigüidade da

linguagem. [5] focaliza em seu trabalho a análise da informação presente nas redes sociais, a extração de conhecimento a partir de grafos e a visualização de fatos decorrentes dessa análise. Para isto, foi conduzido um estudo de caso com base na análise da relação de co-autorias entre pesquisadores, medida a partir da publicação de artigos científicos. A extração de conhecimento foi realizada a partir da aplicação de técnicas de Mineração de Dados, sendo que os dados originais estavam no formato XML, os quais são textuais.

Em [12] foi apresentado um sistema para extrair Regras de Associação automaticamente de páginas de notícias da Internet que tratam de doenças relacionadas a pássaros. O sistema depende da característica da palavra para extrair Regras de Associação.

As citações apresentadas anteriormente confirmam a importância da proposta deste estudo, na medida em que mostram a necessidade da utilização de novas técnicas e ferramentas para trabalhar com textos. No entanto, esta pesquisa mostra um diferencial com relação aos demais, uma vez que aborda, tanto de maneira conceitual, quanto de maneira prática a aplicação do processo de KDT, mostrando a importância dessa abordagem na área de segurança pública.

6 Considerações Finais

Devido o crescente aumento de textos presentes na maioria das instituições, cada vez mais se necessitam de técnicas que possam analisar essas informações de forma a gerar novos conhecimentos para a tomada de decisão. Descobrir conhecimento em textos é bem mais complexo que realizar essa busca em dados, sendo ainda uma área de pesquisa considerada recente no contexto da computação.

Este trabalho analisou o processo de KDT, permitindo analisar as principais características das etapas de cada processo. A principal contribuição deste artigo foi extrair conhecimentos de textos da área de segurança pública com a finalidade de analisar as peculiaridades do crime de Estupro ocorridos no Estado do Pará a partir da aplicação de Regras de Associação a fim de entender e explorar padrões de comportamento dos textos.

Outra contribuição desta pesquisa foi visualizar um contexto, no caso a segurança pública, em que a aplicação de técnicas de KDT pudessem ser aplicadas de modo a permitir uma análise e um entendimento mais profundo acerca das especificações desse crime, podendo ser úteis para os tomadores de decisão na área de segurança pública a fim de tomarem medidas que reduzam a quantidades de ocorrências desse crime, bem como viabilizar a extensão do uso dessas técnicas para aplicações em outros tipos de crimes.

Dentre os resultados encontrados, pode observar uma importante característica do perfil do agressor que pratica o delito estupro, que muitas vezes é conhecido da família e pratica o delito na própria casa da vítima.

Agradecimentos

A FAPESPA - Fundação de Amparo à Pesquisa do Estado do Pará - o apoio financeiro, sob forma de bolsa de mestrado, a um dos autores deste artigo.

Referências

1. Aranha, Cristian; Passos, Emmanuel. Automatic NLP for Competitive Intelligence; In: Prado, H. A; Ferneda, E. (ed.). Emerging Technologies of Text Mining: Techniques and Applications. Hershey, New York: Information Science Reference, 2007, p. 54-76.
2. Ahonen, H.; Heinonen, O.; Mika, Klemettinen; M.; Verkamo, A. I.Applying Data Mining Techniques in Text Analysis *In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD'97*), Newport Beach, California, USA, August 1997.
3. Chen, H. Knowledge Management Systems: A Text Mining Perspective. University of Arizona (Knowledge Computing Corporation), Tucson, Arizona, 2001.
4. Feldman, R. Dagan, I. Knowledge Discovery in Textual Databases (KDT). *In Proceeding. of the 1st International Conference on Knowledge Discovery and Data Mining (KDD-95)*, 112 – 117. AAAI / MIT Press: Menlo Park, CA, 1995.
5. Freitas, C.; Nedel, L.; Galante, R.; Lamb, L.; Spritzer, A.; Fujii, S., Oliveira, J. P.; Araújo, R.; Moro, M. Extração de Conhecimento e Análise Visual de Redes Sociais. XXVIII SEMISH - Seminário Integrado de Software e Hardware. Belém, Pará. 2008.
6. Giudici, P. Applied Data Mining: Statistical Methods for Business and Industry. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England. 2003.
7. Gospodnetic, O.; Hatcher, E. Lucene In Action: A guide to the Java search engine. Manning Publications Co. 2005.
8. Kao, A. and Poteet, S.R. (eds.) Natural Language Processing and Text Mining, Springer, New York, 2006.
9. Lima, V.; Nunes, M. G.; Vieira, R. Desafios do Processamento de Línguas Naturais. XXVII SEMISH - Seminário Integrado de Software e Hardware. Rio de Janeiro, RJ. 2007.
10. Magalhães, M.N; Lima, A.C.P. Noções de Probabilidade e Estatística. 6.Edição , São Paulo: Edusp, 2004.
11. Mathiak, B. e Eckstein, S. (2004). Five Steps to Text Mining in Biomedical Literature. *In: Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics*, held in Conjunction with ECML/PKDD in Pisa, Italy.
12. Mahgoub, Hany; Dietmar Rosner; Nabil Ismail; Fawzy Torkey. A Text Mining Technique Using Association Rules Extraction, International Journal of Computational Intelligence, 2007, v. 4, n. 1, ISSN 1304-2386.
13. Meira Jr, W.; Ferreira, R.; Guedes, D. Escalabilidade e Eficiência em Mineração de Dados de Aplicações Internet. XXVII SEMISH - Seminário Integrado de Software e Hardware. Rio de Janeiro, RJ. 2007.
14. Silva Filho, L. A. ; Betini, R. C. ; Ribeiro, T. V. B. ; Moreira, P. D. O. ; Santos, F. H. Descoberta de Conhecimento no Apoio à Tomada de Decisão na Segurança Pública. *In: CLEI - Conferencia Latinoamericana de Informática*, 2007, Sán José – Costa Rica.
15. Silva Filho, L. A. ; Santos, F. H. ; Moreira, P. D. O. ; Betini, R. C.; Dias, M. M. Descoberta de Conhecimento na Segurança Pública Utilizando Mineração de Dados. In: Edson Marcos Leal Soares Ramos; Silvia dos Santos de Almeida; Adrilayne dos Reis Araújo. (Org.). Segurança Pública Uma Abordagem Estatística e Computacional. Belém: Editora da UFPA, 2008, v. v.2, p. 11-22.
16. Tan, A. Text mining: the state of the art and the challenges. In: Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases - PAKDD'99, p. 65- 77, Beijing, April 1999.