Revista de Ciências Exatas e Tecnologia
Vol. III, Nº. 3, Ano 2008

MINERAÇÃO DE TEXTOS

Text mining

Eliana Cristina Nogueira Barion
Faculdade Anhanguera de Matão
elianabarion@gmail.com

Decio Lago
Faculdade Anhanguera de Matão
d.lago@unianhanguera.edu.br

RESUMO

A Mineração de Textos, também conhecida como Descoberta de Conhecimento em Textos (Knowledge Discovered in Texts – KDT), refere-se ao processo de extração de informação útil (conhecimento) em documentos de textos não-estruturados. O KDT utiliza abordagens já consagradas das áreas de Recuperação de Informação, Processamento de Linguagem Natural e Descoberta de Conhecimento em Banco de Dados. Pelo fato de muitas informações (mais de 80%) estarem armazenadas em formato texto, acredita-se que as técnicas de mineração de textos possuam um grande valor comercial. Este artigo apresenta as técnicas utilizadas para mineração de informação em textos, explicando a funcionalidade e importância de cada uma delas; contudo, a escolha da técnica a ser utilizada depende do objetivo da aplicação.

Palavras-Chave: Mineração de textos, descoberta de conhecimento em banco de dados, descoberta de conhecimento em textos.

ABSTRACT

The Text Mining is also known as Knowledge Discovered in Texts (KDT) – refers to the process of useful information extraction (knowledge) in non-structured texts. The KDT uses approaches once renowned in the areas of Information Recovery, Natural Language Process and Discovery of Knowledge in Database. Due to the fact of much information (more than 80%) are filed in text format, it is believed that the mining techniques have a great commercial value. This article presents the technique used for mining of information on texts, explaining its functionability and their importance as well; however, the choice of technique to be used depends on the application goal.

Keywords: Text mining, knowledge discovery in database, knowledge discovered in texts.

Anhanguera Educacional S.A.

Correspondência/Contato
Alameda Maria Tereza, 2000
Valinhos, São Paulo
CEP. 13.278-181
rc.ipade@unianhanguera.edu.br

Coordenação
Instituto de Pesquisas Aplicadas e Desenvolvimento Educacional - IPADE

Informe Técnico
Recebido em: 14/7/2008
Avaliado em: 27/10/2008

Publicação: 8 de dezembro de 2008

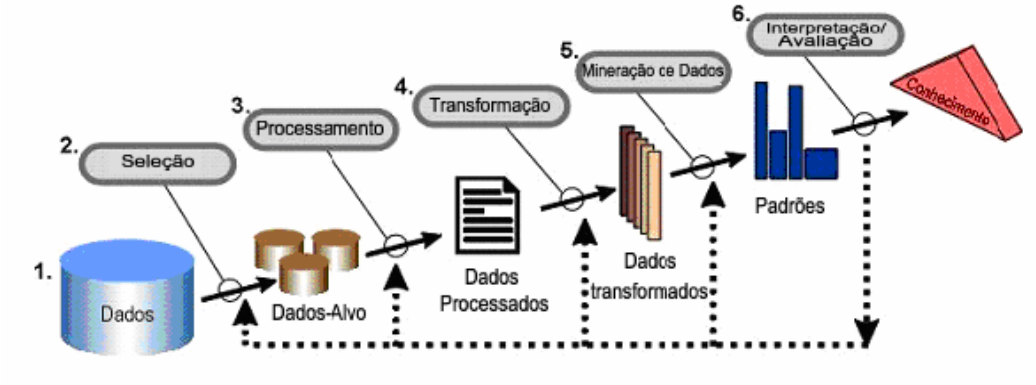
1. INTRODUÇÃO

Com o avanço da informatização surge, cada vez mais, um grande volume de informação armazenado em banco de dados. Estes dados, na maioria das vezes, incluem informações valiosas, por exemplo, tendências e padrões que poderiam ser usados para auxiliar nas tomadas de decisões dentro das empresas. Entretanto, apesar das grandes evoluções surgidas nos Sistemas Gerenciadores de Banco de Dados, torna-se impossível extrair conhecimento para tomadas de decisões a partir destas bases de dados.

Muitas técnicas foram estudadas e desenvolvidas com o objetivo de auxiliarem na extração de informações importantes, implícitas nas bases de dados, dando origem à chamada Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovered in Databases* – KDD).

Conforme mostrado na Figura 1, as etapas pertencentes ao KDD são:

1. Dados: O KDD se baseia no armazenamento dos dados de forma estruturada;
2. Seleção de Dados: Após ter definido o domínio sobre o qual se pretende executar o processo de descoberta, a próxima etapa é selecionar e coletar o conjunto de dados ou variáveis necessárias;
3. Processamento: Esta etapa é também conhecida com pré-processamento visando eliminar os dados que não se adequam às informações, com base nos algoritmos, ou seja, dados incompletos, problemas de definição de tipos, eliminação de tuplas repetidas, etc.;
4. Transformação: Nesta etapa os dados deverão ser armazenados adequadamente para facilitar na utilização das técnicas de mineração de dados;
5. Mineração de Dados: A atividade de descoberta do conhecimento é onde são processados os algoritmos de aprendizado de máquina e de reconhecimento de padrões. A maioria dos métodos de *Data Mining* são baseados em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação, clusterização, modelos gráficos;
6. Interpretação/Avaliação: Nesta etapa final, os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas, porém devem ser apresentadas de forma que o usuário possa entender e interpretar os resultados obtidos.



Fonte: CORRÊA (2003).

Figura 1. Etapas do KDD.

No entanto, a aplicação destas técnicas, pode-se dar a partir dos dados já estruturados.

Segundo Tan (1999), 80% das informações de uma companhia estão contidas em documentos textuais. Chen (2001) afirma ainda que 80% do conteúdo on-line estão em formato texto. A partir destas afirmações, conclui-se que apenas 20% das informações são usadas para manipulação de tomada de decisão dentro das empresas.

A mineração de textos surge, então, da necessidade de se descobrir, de forma automática, informações (padrões e anomalias) em textos. Mineração de textos é um conjunto de métodos usados para navegar, organizar, achar e descobrir informações em bases de textos. Pode ser vista como uma extensão da área de *Data Mining*, focada na análise de textos. É também chamada de *Text Data Mining*, *Knowledge Discovery in Texts* (KDT). Segundo Passos (2006), a mineração de textos é um campo multidisciplinar que inclui conhecimentos de áreas como Informática, Estatística, Lingüística e Ciência Cognitiva.

Com base no conhecimento extraído dessas ciências, a mineração de textos define técnicas de extração de padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente, para objetivos específicos. Inspirado pelo *data mining* ou mineração de dados, que procura descobrir padrões emergentes de banco de dados estruturados, a mineração de textos pretende extrair conhecimentos úteis de dados não estruturados ou semi-estruturados.

A descoberta de conhecimento em textos conta, basicamente, com duas etapas: a primeira etapa consiste no tratamento do texto a fim de convertê-lo para uma forma estruturada; a segunda etapa consiste na aplicação da mineração para a descoberta do conhecimento (CORRÊA, 2003).

As técnicas de Processamento de Linguagem Natural (PLN) e Recuperação de Informação são aplicadas na primeira etapa e a Descoberta de Conhecimento em Banco de Dados é aplicada na segunda etapa.

2. DESCOBERTA DE CONHECIMENTO EM TEXTOS (KDT)

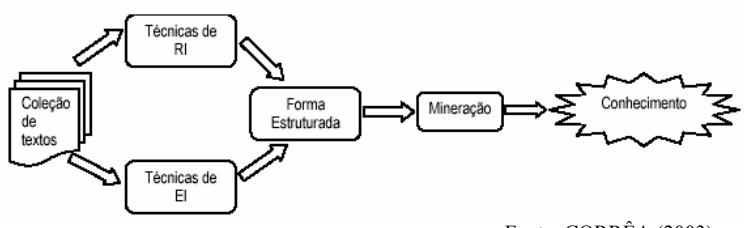
O KDT engloba técnicas e ferramentas inteligentes e automáticas que auxiliam na análise de grandes volumes de dados com o intuito de garimpar conhecimento útil, auxiliando nas tomadas de decisões, possibilitando a descoberta de estratégias organizacionais.

É possível aplicar as técnicas de descoberta de conhecimento em informações textuais. No entanto, os bancos de dados textuais apresentam-se desestruturados, impossibilitando a aplicação das técnicas utilizadas em bancos de dados estruturados. Técnicas específicas para tratamento de textos devem ser utilizadas a fim de se obter conhecimentos implícitos em banco de dados textuais.

As etapas da mineração de textos serão descritas nos tópicos a seguir.

3. ETAPAS DA MINERAÇÃO DE TEXTOS

Como se pode observar na Figura 2, as técnicas de Recuperação de Informação ou as técnicas de Extração de Informação são aplicadas sobre uma coleção de textos para que se obtenha uma forma estruturada; e, a partir dos dados já estruturados, são aplicadas as técnicas de Mineração de Dados para que se obtenha o conhecimento.



Fonte: CORRÊA (2003).

Figura 2. Processo de Mineração de Textos.

3.1. Recuperação da informação

Segundo Salton e McGill (1983), os processos para recuperação de informação necessitam de técnicas que agilizam o armazenamento e acesso aos dados. Estas técnicas en-

volvem a atribuição de termos apropriados e identificadores para representar o conteúdo dos documentos na coleção. Esta tarefa, conhecida como indexação, pode ser feita automaticamente ou manualmente.

A Recuperação da Informação é feita através de uma entrada do usuário, ou seja, através de uma consulta para que os documentos relevantes sejam encontrados. Os processos de Recuperação de Informação geralmente se baseiam em Buscas por Palavra-Chave ou Busca por Similaridade (KAMBER, 2001).

No processo baseado por Busca por Palavra-Chave um documento é representado como um conjunto de termos e pode ser identificado por palavras-chave.

O processo baseado por Busca por Similaridade pode ser representado pelo modelo vetorial, criado por Salton e McGill (1983). O modelo vetorial representa as consultas e os documentos como vetores de termos. O cálculo de similaridade entre o vetor que representa a consulta e o vetor de documentos é o vetor resultado para uma consulta. A relevância dos termos, tanto para as consultas quanto para os documentos, é quantificada pelos pesos relacionados a cada termo do vetor.

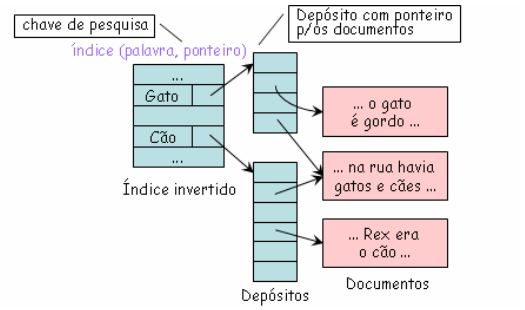
A seguir, será representado o processo de indexação e os cálculos utilizados para a indicação de relevância dos termos para os textos.

3.2. Processo de indexação

A indexação é o processo pelo qual as palavras contidas no texto são armazenadas em uma estrutura de índices para viabilizar a pesquisa de documentos através das palavras que ele contém (SALTON; MCGILL, 1983).

Arquivos invertidos são tradicionalmente usados para a implementação de índices lexicográficos, ou seja, de índices ordenados. Aplicado ao contexto de pesquisas por frases, um arquivo invertido pode ser visto como uma lista ordenada de palavras-chave contendo, para cada palavra, um apontador para cada um dos documentos em que a palavra ocorre, juntamente com a posição da palavra nesse documento. Os índices invertidos são usados para melhorar o desempenho e a funcionalidade das buscas (ROCHA, 2002).

Na Figura 3 é ilustrado o comportamento de índices invertidos, mostrando que os termos ou palavras chaves são extraídos dos textos e ficam armazenados juntamente com as referências para os respectivos documentos.



Fonte: ROCHA (2002).

Figura 3. Estrutura de Índices Invertidos.

O processo de indexação é composto de: Análise Léxica, Remoção de *Stop-Words*, *Stemming*, Seleção dos termos-índice, Determinação de Pesos e Criação de Tesauros.

A Análise Léxica é a etapa para converter uma seqüência de caracteres (o texto dos documentos) numa seqüência de palavras que serão as palavras candidatas a serem termos do índice. O analisador léxico separa o alfabeto de entrada em caracteres de palavras (a-z) e separadores de palavras (espaço, nova linha, etc.).

O processo de Remoção de *Stop-Words* é utilizado para remover um conjunto de palavras que aparecem com muita freqüência no texto. Estas palavras, chamadas de *Stop-Words*, geralmente são preposições, artigos, conjunções, alguns verbos, nomes, adjetivos e advérbios. Para isto, deve ser criada uma lista, denominada *Stop-List*, no idioma em que se está trabalhando, contendo estas palavras consideradas irrelevantes. Este processo faz-se necessário para retirar do texto palavras que não tem nenhuma importância, diminuindo assim o tamanho das estruturas de indexação e facilitando a mineração.

O processo de *Stemming* é utilizado para remover todas as variações de palavras permanecendo somente a raiz. Estas variações são prefixos, sufixos que são removidos das palavras melhorando o armazenamento por eliminar a quantidade de termos a serem armazenados. Por exemplo, a palavra *computer* pode conter muitas variações (*computers*, *computing*, *computation*) que têm semânticas similares, mas que relacionam-se ao mesmo conceito. Plurais e gerúndios (ing) também fazem parte destas variações a serem removidas.

A remoção de *Stop-Words* e *Stemming* devem ser retiradas antes do processo de indexação, melhorando o tamanho das estruturas.

A Seleção de termos-índice é usada para determinar quais palavras ou radicais serão usados como elementos de indexação. As palavras-chave selecionadas são as que

possuem os maiores valores para o peso. Os substantivos freqüentemente transmitem mais significados semânticos. É muito comum ter-se o agrupamento de substantivos que aparecem próximos no texto, combinando-os em um componente de indexação único.

Por exemplo, a frase: "A destruição das florestas tropicais da Amazônia" primeiramente é normalizada para letras minúsculas, depois é feito a Remoção de palavras de ligação (*Stop Words*): "*destruição florestas tropicais amazônia*". A seguir é feito a Remoção de sufixo (*Stemming*): "destru florest tropic amazon".

Na etapa de Determinação de Pesos são utilizadas medidas de freqüência relativas, capazes de identificar termos que ocorrem com substancial freqüência em alguns documentos de uma coleção, mas com baixa freqüência na coleção toda. Dentre as medidas de freqüência relativa, destaca-se o peso do termo. O peso do termo consiste em assumir que sua importância é proporcional à freqüência de ocorrência de cada termo k em cada documento i e inversamente proporcional ao número de documentos para os quais o termo é encontrado (CORRÊA, 2003)

Usando a expressão proposta por Salton e McGill (1983), tem-se o seguinte cálculo para determinar o peso do termo-índice:

$$\text{Peso} = \frac{FREQIK * \log_2(n)}{\text{DOCFREQ}_K} + 1 \qquad (1)$$

Onde *FREQIK* é o número de vezes em que o termo aparece no texto e DOCFREQ_K é o número de documentos da coleção. É importante ressaltar que nem todos os termos são igualmente úteis para representar o conteúdo do documento. Geralmente termos menos freqüentes permitem identificar um conjunto mais restrito de documentos.

Exemplo 1: Termo = "*mining*" Nº. de Documentos = 1 Freqüência = 7 vezes *Peso* = 7 * *log2*(7/1) + 1 = 21	Exemplo 2: Termo = "*mining*" Nº. de Documentos = 10 Freqüência = 7 vezes *Peso* = 7 * *log2*(7/10) + 1 = 5,35

Essa fórmula leva em consideração que termos que ocorrem com substancial freqüência em alguns documentos são muito mais importantes que termos que ocorrem com grande freqüência na coleção toda.

Etapa de Criação de Tesauros - Segundo Salton e McGill (1983), além do processo de indexação alguns refinamentos são sugeridos, consistindo de associações entre termos, conhecidas como classes Tesauros.

Um tesauro pode ser definido como um vocabulário controlado que representa hierarquias, relações de equivalência, pertinência e associações entre os termos, com o objetivo de auxiliar o usuário potencial a encontrar a informação de que necessita com a menor margem de erro possível (COLEPÍCOLO, 2004).

Um determinado tesauro pode pertencer a um domínio de conhecimento ou ainda pode ser genérico para cada língua. É representado como um grafo onde cada nó representa a um termo relacionado a outros termos.

Segundo Baeza-Yates (1999), para se criar um tesauro é necessário calcular a similaridade entre os pares de termos encontrados durante o processo de indexação.

A fórmula utilizada para o cálculo de similaridade, sugerida por Salton e McGill (1983), baseia-se num vetor de documentos contendo os pesos associados de cada termo para o referido documento.

Exemplo do uso do cálculo de similaridade proposto por Salton e McGill (1983): Supondo que se tem os seguintes termos associados aos documentos [d1..d4] e se deseja calcular a similaridade entre os termos *clusterization* e *categorization*:

Matriz termo a termo

Vetores de Documentos		Mining	Discovery	Algorithms	Categorization	Clusterization	Dictionary
	D1	3	4	0	0	0	0
	D2	2	3	2	0	0	1
	D3	0	0	3	5	7	1
	D4	1	4	1	0	1	0

Fonte: CORRÊA (2003)

Figura 4. Matriz de Vetores de Documentos.

Uma matriz de vetores de documentos, contendo suas palavras chaves e os pesos associados são mostrados na Figura 4.

Veja a fórmula do cálculo de similaridade mostrado na equação 1, segundo Salton e McGill (1983):

$$SIMILAR(TERMOk, TERMOh) = \frac{\sum_{i=1}^{n} w_{ik} w_{ih}}{\sum_{i=1}^{n} (w_{ik})^2 + \sum_{i=1}^{n} (w_{ih})^2 - \sum_{i=1}^{n} w_{ik} w_{ih}} \quad (2)$$

$$\textbf{\textit{categorization,clusterization}} = \underline{(0*0+0*0+5*7+0*1)} = \underline{35} = 0{,}87$$

$$(0^2+0^2+0^2+0^2+5^2+7^2+0^2+1^2)-(0*0+0*0+5*7+0*1)\ 40$$

Cada linha da matriz representa um vetor de documentos e as colunas representam os termos associados aos documentos. Segundo Salton e McGill (1983), dados vetores de termos na forma $TERMO_j = (W_{1J}, W_{t2}j,...W_{nj})$, onde w_{ij} indica o $TERMO_j$ no documento i e assumindo n documentos na coleção.

A partir deste cálculo, tem-se que o grau de similaridade entre os termos *categorization* e *clusterization* é de 0,87.

O resultado do cálculo anterior para cada par de palavras-chave gera a matriz de similaridade, como é ilustrado na Figura 5:

Termos de Indexação de Documentos – Matriz de Similaridade								
Valores de Indexação de Documentos		*Mining*	*Discovery*	*Algorithms*	*Categorization*	*Clusterization*	*Dictionary*	*databases*
	Mining	-	0,66	0,22	0	0	0,14	0,26
	Discovery	0,66	-	0,22	0	0	0,07	0,32
	Algorithms	0,22	0,22	-	0,62	0,52	0,45	0,58
	Categorization	0	0	0,62	-	0,87	0,23	0,68
	Clusterization	0	0	0,52	0,87	-	0,15	0,68
	Dictionary	0,14	0,07	0,45	0,23	0,15	-	0,18
	databases	0,26	0,32	0,58	0,68	0,68	0,18	-

Fonte: CORRÊA (2003)

Figura 5. Matriz de Similaridade.

Estabelecendo-se um coeficiente de similaridade de 0,6 (k=0,6), gera-se uma matriz binária, conforme mostrado na Figura 6, atribuindo o valor 1 (um) para graus de similaridades superiores a 0,6 e valor zero para similaridades inferiores a este valor.

Termos de Indexação de Documentos – Matriz Binária								
Valores de Indexação de Documentos		*Mining*	*Discovery*	*Algorithms*	*Categorization*	*Clusterization*	*Dictionary*	*databases*
	Mining	-	1	0	0	0	0	0
	Discovery	1	-	0	0	0	0	0
	Algorithms	0	0	-	1	0	0	0
	Categorization	0	0	1	-	1	0	1
	Clusterization	0	0	0	1	-	0	1
	Dictionary	0	0	0	0	0	-	0
	databases	0	0	0	1	1	0	-

Fonte: CORRÊA (2003)

Figura 6. Matriz Binária para k=0,6.

Baseando-se na matriz binária gerada, usa-se o método de classificação automática (*single-link*) para construir classes de termos similares (equivalentes a classes tesauros) (SALTON; MCGILL, 1983). O método utilizado agrupa em uma classe comum todos os termos cujos coeficientes de similaridade estejam dentro do padrão estabelecido (superior a 0,6).

Baseado na matriz binária aplica-se o modelo de classificação automática. Veja a Tabela 1:

Tabela 1- Modelo de classificação automática.

Modelo de Classificação Automática	
Termo Original	**Termos Similaridade**
Mining	*Discorery*
Discovery	*Mining*
Algorithms	*Categorization*
Categorization	*Algorithms, clusterization, databases*
Clusterization	*Categorization, databases*
Databases	*Categorization, clusterization*

Fonte: CORRÊA (2003)

A partir do exemplo apresentado, o usuário seleciona o termo desejado e o sistema busca todos os termos relacionados. Neste caso, se o usuário procura por "*clusterization*", o sistema procura a classe onde está a palavra "*clusterization*", detecta também a palavra "*categorization*" e "*databases*".

Estratégias de busca

A partir das estruturas de dados e da consulta formulada, recupera-se uma lista de documentos considerados relevantes. Existem três modelos de recuperação para os Sistemas de Recuperação de Informação. São eles: modelo Booleano, Vetorial e Probabilístico.

O modelo Booleano é baseado na álgebra booleana e é considerado o modelo mais simples de recuperação. Considera uma consulta como uma expressão booleana convencional formada pelos conectivos lógicos AND, OR e NOT.

AND = Intersecção: documentos retornados devem ter ambas as palavras

OR = União: retorna documentos que tenham ambas as palavras independente de estarem no mesmo documento.

NOT = Negação: retorna os documentos que tenham X, mas que não tenham Y.

O modelo Vetorial representa documentos e consultas como vetores de termos. A consulta é construída baseada num ângulo de similaridade entre o vetor que representa o documento e o vetor que representa a consulta.

O modelo probabilístico descreve documentos considerando pesos binários que representam a presença ou ausência de termos.

Medidas de eficácia

As Medidas de Eficácia verificam o grau de satisfação da resposta para uma determinada consulta. Os critérios mais comumente utilizados na literatura são precisão e abrangência, ou, *precision* e *recall*.

Precision é a porcentagem de respostas relevantes efetivamente recuperadas numa consulta em relação ao total de respostas obtidas. *Recall* é a porcentagem de todas as respostas (documentos) relevantes que são efetivamente recuperados por uma consulta em relação ao total de respostas relevantes previstas (CORRÊA, 2003).

3.3. Extração de informação

A extração de informação é usada na área de Processamento de Linguagem Natural (PLN) com o objetivo de transformar dados semi-estruturados ou desestruturados (textos) em dados estruturados que serão armazenados em um banco de dados. Uma vez

estruturados estes dados podem ser usados para processos tradicionais de descoberta de conhecimento (CORRÊA, 2003).

O processo de extração de informação identifica palavras dentro de conceitos específicos e ainda contém um processo de transformação que modifica a informação extraída em um formato compatível com um banco de dados.

Processo de extração

O processo de Extração de Informação, a partir de Processamento de Linguagem Natural, é usado com o objetivo de transformar dados desestruturados (textos) ou semi-estruturados em dados estruturados a fim de que sejam armazenados em um banco de dados para obtenção do conhecimento.

Este processo de extração deve ser feito sobre um tipo de domínio com as informações pré-definias do que se deseja encontrar no texto. Este domínio é chamado de *Slots*. Fazendo-se uma analogia, os *slots* podem ser comparados a atributo-valor nos bancos de dados tradicionais; devendo conter as informações que se deseja extrair do texto. Por exemplo, em textos sobre transmissões de doenças, os *slots* poderiam ser preenchidos com as informações: nome da doença, meios de transmissão, origem da doença, etc. Estas lacunas a serem preenchidas com as informações extraídas do texto são denominadas *templates*.

Segundo Corrêa (2003), os Sistemas de Extração de Informação não tentam interpretar o texto em todas as partes do documento de entrada, mas sim analisar partes do texto que possuam informações relevantes ao domínio específico. O sistema de extração de Informação trata uma coleção de textos, isolando fragmentos de texto para a extração de informações relevantes e armazena as informações extraídas na forma tabular.

O processo de Extração de Informação, ilustrado pela Figura 7, envolve a criação dos *slots* contendo as informações desejadas do domínio até a geração das *templates*.

As *templates* podem ser definidas como se fossem documentos com lacunas a serem preenchidas com as informações relevantes extraídas dos textos.

Fonte: CORRÊA (2003).

Figura 7. Extração de Informação.

Com base nos *slots* criados o texto é analisado pelo analisador léxico preenchendo estes *slots*. O texto todo é marcado com *tags* SGML[1] (*Standard Generation Markup Language*) para que o texto seja delimitado pelas informações a serem extraídas, obtendo-se, desta forma, um texto semi-estruturado.

Estas informações são extraídas e submetidas a tratamentos para serem armazenadas na forma tabular ou formatadas em um texto em linguagem natural. Utiliza-se um extrator em cada documento para que se crie uma coleção de registros estruturados onde são aplicadas as técnicas de Mineração de Dados para que se descubram relacionamentos interessantes.

Marcação POS - Part-Of-Speech

As principais técnicas de Extração de Informação reconhecem as estruturas de um texto através da análise de *tags* (marcas que possam ser identificadas no texto). A marcação POS automaticamente atribui marcas de discurso para as palavras no contexto. Estas marcas de discurso definem as categorias morfossintáticas das palavras. Cowie e Lehnert (1996) citam como exemplos o reconhecimento de nomes próprios (como nomes de pessoas, companhias, etc.), por começarem com uma letra maiúscula e por, geralmente, virem próximos de termos como "Senhor", "limitada", etc.

A ambigüidade léxica é a principal dificuldade para esta técnica de marcação. Por exemplo, quando temos palavras que podem ser tanto verbo como substantivo.

Técnicas de Rapier. Algumas técnicas são utilizadas pelos *taggers* para marcação de texto. Um exemplo é a técnica de RAPIER utilizada sobre alguns documentos com *slots* preenchidos (chamados de conjuntos de treinamento) para que se adquira a base de conhecimento das regras de extração e então possa ser utilizado em novos documentos.

[1] Sistema de *tags* do qual se pode definir linguagens de marcação para documentos.

A Figura 8 referencia-se a um anúncio de emprego e mostra um exemplo de *slot* preenchido:

title: Sênior Software Developer
city: Austin
language: Perl, C, Javascript, Java
platform: NT, Windows
application: Oracle, Informix, Sybase
required years of experience: 5

Fonte: CORRÊA (2003)

Figura 8. Exemplo de slots preenchidos.

O RAPIER é um sistema de aprendizado de máquina para induzir regras na extração de informação de textos em linguagem natural. Esta técnica aprende regras a partir dos *slots* preenchidos (CORRÊA , 2003).

Perl ∈ *language*
C ∈ *language*
Javascript ∈ *language*
NT ∈ *platform*
Windows ∈ *platform*
Oracle ∈ *application*

Fonte: CORRÊA (2003)

Figura 9. Técnica Rapier: Exemplo de regras.

Veja na Figura 9 algumas regras aprendidas a partir dos *slots* preenchidos O documento gerado com as *tags* é apresentado na Figura 10.

<DOCUMENT>
Internet/N Provider/N using/V cutting/ADJ edge/N web/N technology/N in/PR **<CITY>** Austin **</CITY>** is/V accepting/V applications/N for/PR a/DET **<TITLE>** Senior_Software_Developer **</TITLE>**. The candidate/N must/V have/V **<YEARS_EXPERIENCE>** 5_years **</YEARS_EXPERIENCE>** of/PR software/N development/N and experience with **<APPLICATION>** Oracle, Sybase **</APPLICATION>**.
…
</DOCUMENT>

Fonte: CORRÊA (2003).

Figura 10. Documento com as *tags*.

3.4. Técnicas de mineração de textos

A partir de todo o processo para estruturação de dados, técnicas de mineração de dados são aplicadas sobre o banco de dados gerado. A seguir são apresentadas rapidamente cada fase da etapa de mineração, visto que estas técnicas já são conhecidas pelo processo de *Data Mining*.

Técnicas de associação

A extração de regras de associação é uma técnica de *Data mining* que gera regras do tipo "Se X Então Y" a partir de um banco de dados de transações, onde X e Y são conjuntos de itens que co-ocorrem em várias transações (SANTOS, 2002).

A motivação por trás de um algoritmo que gera regras de associação relaciona-se com a grande quantidade de aplicações possíveis para estas regras. Questões a respeito das características de consumo sempre foram analisadas com o objetivo de maximizar não apenas a quantidade de vendas, mas também a quantidade de vendas de certos produtos. Na mineração de regras de associação é comum encontrar questões como Pichiliani (2008):

> O que caracteriza quem compra o produto X?
> O que é tipicamente comprado juntamente com o produto Y?
> Que pares de produtos são comprados em conjunto?
> Quem compra isso e aquilo, compra a seguir o quê?

O algoritmo indica a correlação dos produtos e, a partir do conhecimento do fato, uma ação é tomada. Com um pouco mais de formalidade, pode-se definir as regras de associação da seguinte forma:

X **=>** Y, onde X e Y são conjuntos que podem conter um ou mais elementos. O conjunto total de transações pode ser chamado de chamado de conjunto T.

Esta técnica é bastante utilizada em mineração de textos, com o objetivo de descobrir as associações existentes entre termos e categorias de documentos.

As tarefas de associação utilizam algoritmos específicos, dentre eles destaca-se o algoritmo APriori, utilizado para encontrar associações relevantes entre itens de dados.

O algoritmo APriori quando é aplicado em algum texto encontra conjuntos freqüentes de palavras nos documentos do conjunto de treinamento. As regras utilizadas são do tipo X => Y, onde X é um conjunto de palavras e Y é uma categoria. Para

cada categoria podem ser aplicadas diferentes regras e desta forma, o classificador obtido é um conjunto de regras de cada categoria. Maiores detalhes sobre algoritmos A-Priori podem ser encontrados em (CORRÊA, 2003).

Sumarização

O processo de sumarização seleciona as informações mais importantes do texto, tornando a descrição mais compacta, mas mantendo a mesma informação. É uma técnica bastante utilizada em mineração de textos com o intuito de identificar palavras ou frases mais importantes dos documentos.

A sumarização tem por objetivo produzir uma lista de sentenças do documento de origem resumindo o conteúdo deste documento, reduzindo seu volume, mas mantendo a mesma informação.

Clusterização

As técnicas de clusterização são usadas para agrupar um conjunto de dados considerados similares em *clusters* ou grupos. A construção de tesauros, comentadas anteriormente neste artigo, também é obtida a partir de uma matriz de similaridade.

A importância do uso das técnicas de clusterização na mineração de texto é que se extraindo a hierarquia de textos em linguagem natural, os termos adjacentes ou as relações sintáticas entre termos carregam um considerável poder descritivo para inferir a semântica de uma hierarquia de conceitos relacionados a esses termos (MOURA, 2004).

Classificação/Categorização

A classificação ou categorização é um processo que visa a identificação de tópicos principais em um documento e a sua associação baseando-se em um algoritmo pré-definido, construído a partir de um conjunto de treinamento definido por pessoas experientes no assunto envolvido.

Este algoritmo analisa todos os exemplos de documentos, aprende as regras e as armazena em uma Base de Conhecimento. Os documentos a serem classificados passam por um Categorizador, o qual, baseado nas regras previamente inseridas na Base de Conhecimento, estabelece a qual classe pertence cada documento (CORRÊA, 2003).

4. CONSIDERAÇÕES FINAIS

A mineração de textos ou Descoberta de Conhecimento em Textos, difere de um mecanismo de busca. Na busca, o usuário já tem o conhecimento do que deseja encontrar enquanto que a mineração de textos auxilia o usuário na descoberta de informações desconhecidas.

Como muitas informações estão armazenadas em forma de texto (mais de 80%), as técnicas de mineração de textos são muito importantes para a recuperação do conhecimento implícito nestes documentos. Sendo assim, muitos estudos ainda estão sendo feitos a fim de aprimorar e descobrir novas técnicas para a descoberta de conhecimentos em bases textuais.

Este artigo apresentou as Técnicas de Processamento de Linguagem Natural, Recuperação de Informação e Extração da Informação que são aplicadas para estruturar a base de dados e as técnicas de Descoberta de Conhecimento em Banco de Dados aplicadas a fim de descobrir informações de tendências e padrões.

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO NETO, B. **Modern information retrieval**. Addison-Wesley, 1999.

CHEN, H., **Knowledge management systems**: a text mining perspective. University of Arizona (Knowledge Computing Corporation), Tucson, Arizona, 2001.

COLEPÍCOLO, Eliane; HOLANDA, Adriano J.; RUIZ, Evandro E. S.; WAINER, Jacques; PISA, Ivan T., **MeSH**: de cabeçalho de assunto a tesauro. USP, UNIFESP 2004. Disponível em: <http://www.sbis.org.br/cbis/arquivos/994.pdf>. Acesso em: 03 set. 2007.

CORRÊA, Adriana Cristina Giusti. **Recuperação de documentos baseada em Informação Semântica no Ambiente AMMO**. UFSCAR 2003. Disponível em: <http://www.bdtd.ufscar.br/tde_busca/arquivo.php?codArquivo=485>. Acesso em: 23 ago. 2007.

COWIE, J.; LEHNERT, W. InformationExtraction. **Communications of the ACM**, v. 39, n. 1, jan. 1996.

EMBRAPA. **Seleção, classificação e qualificação de documentos**. 2004. Disponível em: <http://www.cnptia.embrapa.br/modules/tinycontent3/content/2004/doc47.pdf>. Acesso em: 18 set. 2007.

KAMBER, M.; HAN, J. **Data mining**: concepts and techniques. Morgan Kaufmann, 2001.

MOURA, M. F. **Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos**. Campinas: Embrapa Informática Agropecuária, 2004 (Embrapa Informática Agropecuária. Documentos).

PASSOS, Emmanuel; ARANHA, Christian. **A tecnologia de mineração de textos**. UFSC, 2006. Disponível em: <http://www.inf.ufsc.br/resi/edicao08/Artigo86Tutorial Emmanuel.pdf>. Acesso em: 29 ago. 2007.

PICHILIANI, Mauro. **Data mining na prática**: regras de associação. 2008. Disponível em: <http://imasters.uol.com.br/artigo/7853/sql_server/data_mining_na_pratica_regras_de_associacao>. Acesso em: 10 jul. 2008.

ROCHA, Marcus V.; DA COSTA, Mateus Conrad B.; DOS SANTOS NETO, Pedro de Alcântara. **Busca por Frases em Bancos de Dados Textuais**. UFMG, 2002. Disponível em <http://homepages.dcc.ufmg.br/~nivio/cursos/pa02/seminarios/seminario3/seminari3.html>. Acesso em: 03 set 2007.

SALTON, G.; MCGILL, M. J. **Introdution to modern information retrieval**. Computer Science Series, USA: McGraw-Hill, 1983.

SANTOS, Maria Angela Moscalewski Roveredo. **Extraindo Regras de Associação a partir de Textos**. PUC, 2002. Disponível em: <http://www.ppgia. pucpr.br/teses/DissertacaoPPGIa-MariaRoveredo-062002.pdf>. Acesso em: 03 set. 2007.

TAN, A. H. Text mining: the state of the art and the challenges. In: **Proceddings...**, PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, Beijing, 1999, p. 65-70.

Eliana Cristina Nogueira Barion

Especialista em Sistemas de Informação – ASSER, São Carlos, SP. Mestranda em Ciência da Computação – UFSCar. Professora da Faculdade Anhanguera de Matão.

Decio Lago

Mestre em Desenvolvimento Regional e Meio Ambiente – UNIARA. Especialista em Geoprocessamento – Ngeo – UFSCar. Coordenador do curso Sistemas de Informação da Faculdade Anhanguera de Matão.