A Tecnologia de Mineração de Textos

(Artigo tutorial)

Christian Aranha, Emmanuel Passos
Lab.ICA Elétrica PUC-Rio

Resumo

Mineração de textos, também conhecido como mineração de dados textuais ou descoberta de conhecimento de bases de dados textuais, em geral, se refere ao processo de extração de informações de interesse e padrões não-triviais ou descoberta de conhecimento em documentos de texto não-estruturados. Pode ser visto como uma extensão da mineração de dados ou da descoberta de conhecimento em bases de dados estruturadas.

Como muitas informações (mais de 80%) estão armazenadas em formato texto, acredita-se que as técnicas de mineração de textos possuam um grande valor comercial.

O objetivo deste tutorial é apresentar algumas técnicas de mineração de textos, bem como casos de uso e resultados obtidos

Palavras chave: Mineração de textos, Sistemas de Informação Inteligentes, Mineração de dados

Abstract

Text mining, also known as text data mining or knowledge-discovery in text (KDT), refers generally, to the process of extracting interesting and non-trivial information and knowledge from unstructured text. It can be seen as an extension of data mining or knowledge discovery in structured databases.

As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value.

The objective of this tutorial is present some techniques of text mining, as well as study cases and their results.

Key-words: Text mining, Data minig, Intelligent information sysems

1. Introdução

Mineração de textos, também chamado de mineração de dados textuais ou descoberta de conhecimento de bases de dados textuais é um campo novo e multidisciplinar que inclui conhecimentos de áreas como Informática, Estatística, Linguística e Ciência Cognitiva. Mineração de textos consiste em extrair regularidades, 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 informática é composta pelo conjunto das Ciências da Informação, que inclui a teoria da informação, o processo de cálculo, a análise numérica, os métodos teóricos da representação dos conhecimentos e modelagem dos problemas.

A estatística é uma ciência que utiliza teorias probabilísticas para explicação de eventos, estudos e experimentos. Tem por objetivo obter, organizar e analisar dados, determinar as correlações que apresentem, tirando delas suas consequências para descrição e explicação do que passou e previsão e organização do futuro. Da informática e da estatística surgiram os famosos mecanismos eficientes de busca de informação como Google e Yahoo!

A lingüística é o estudo científico da linguagem humana. Os lingüistas dividem o estudo da linguagem em áreas que são estudadas mais ou menos de forma independente. As divisões mais comuns são: fonética, fonologia, morfologia, sintaxe, pragmática, dentre outras. A preocupação em adequar os modelos à realidade da computação consolidou a Linguística Computacional.

A ciência cognitiva é normalmente definida como o estudo científico da mente ou da inteligência. Quase toda a introdução à ciência cognitiva frisa a sua alta inter-disciplinaridade; é normalmente caracterizada como tomando parte ou colaborando com as disciplinas de psicologia (especialmente através da psicologia cognitiva, linguística, neurociência, inteligência artificial (em particular no ramo de redes neurais) e filosofia (especialmente a filosofia da mente e a filosofia da matemática mas com aplicações na filosofia da ciência).

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.

14. O que é Mineração de Textos?

A tecnologia de mineração de textos vem das técnicas de recuperação de informações, machine learning (que é um ramo do estudo de sistemas de Informação inteligentes que por sua vez é uma das aplicações notáveis da Inteligência Artificial), (ver Barreto [4]) e da descoberta tradicional de informações estruturadas, através do uso de bancos de dados e de procedimentos estatísticos.

Mineração de textos é um conjunto de métodos usados para navegar, organizar, achar e descobrir informação em bases textuais. Pode ser vista como uma extensão da área de Data Mining, focada na análise de textos. Também é chamada de Text Data Mining, Knowledge Discovery in Texts.

"Uma extração não trivial de informações, não explícitas, de grandes bases textuais, previamente desconhecidas, e potencialmentes úteis (ver Feldman, e Hirsh, [8].

3. O que não é Mineração de Textos?

Mineração de Textos (TM) é diferente de um mecanismo de busca. Na busca o usuário já sabe o que quer encontrar. A tecnologia usada em mineração de textos ajuda o usuário a descobrir informações desconhecidas. TM é diferente de análise de constituintes, pois não é necessário formalizar toda a construção sintática do texto. TM é diferente de chatterbot (robôs de conversação), pois não se pretende simular o comportamento humano. TM não é mineração de dados, pois trabalha com textos utiliza algoritmos de mineração de dados além de outros convenientes.

4. Motivação

A mineração de textos surgiu a partir da necessidade de se descobrir, de forma automática, informações (padrões e anomalias) em textos. O uso dessa tecnologia permite recuperar informações, extrair dados, resumir documentos, descobrir padrões, associações e regras e realizar análises qualitativas ou quantitativas em documentos de texto.

O crescimento do armazenamento de dados não estruturados, devido ao avanço da mídia digital, propiciou o desenvolvimento das técnicas de mineração de textos. Normalmente, os documentos onde são aplicadas as técnicas de mineração de textos incluem: emails, textos livres obtidos por resultados de pesquisas, arquivos eletrônicas gerados por editores de textos, páginas da Web, campos textuais em bancos de dados, documentos eletrônicos, digitalizados a partir de papéis.

- 80% do conteúdo online está em formato textual Chen [6].
- 80% das informações armazenadas por uma empresa são também dados não-estruturados (ver Tan [17]).

Segundo Barcelos [3] como as técnicas desenvolvidas para Mineração de Dados foram desenvolvidas para dados estruturados, técnicas específicas para Mineração de Textos têm sido desenvolvidas para processar uma parte importante da informação disponível, que pode ser encontrada na forma de dados não-estruturados.

A conceituação de mineração de textos apresentada em Zanasi [19], nos ajuda a entender melhor a potencialidade desta técnica / ferramenta. Segundo o autor, mineração de textos é o processo de extrair, dirigido pelos dados, conhecimento não conhecido previamente, a partir de fontes textuais (correio, imprensa, transações, websites, newsgroups, fóruns, listas de correspondência, etc.) úteis para tomar decisões cruciais de negócio.

5. Primeiros Passos

– O que precisamos para manipular texto?

Textos fazem parte da categoria de dados não-estruturados. Isso quer dizer que antes de começar a trabalhar com eles é necessário estruturá-los. Isto é, algum procedimento que transforme a seqüência de caracteres em objetos relacionados entre si. A lógica dessa transformação está presente no próprio texto, através de padrões lingüísticos.

– Quais são os pré-requisitos?

Para se entender o processamento de textos é necessário um pouco de conhecimento de áreas como informática, lingüística e ciência cognitiva. E para entender a forma como se processa a sintaxe de uma língua é necessário entender os modelos do significado de uma palavra. E como nós compreendemos essa palavra dentro do texto pode ser bastante útil na resolução de problemas na área.

– Dependência da língua.

Apesar de que todas as línguas que trataremos, são faladas por humanos, e por isso deveriam seguir o mesmo sistema lógico, as variações culturais podem ser bastante drásticas. Um único padrão essencial para entender todas às línguas ainda é utópico. Sendo assim, qualquer sistema de processamento de textos tem parâmetros específicos para cada língua, até os mais genéricos, como o sistema de busca Google, devem armazenar uma lista de palavras sem poder de discriminação de documentos para cada língua.

– Como processamos o texto?

Processar o texto consiste primeiramente em dividí-lo em partes menores aproveitando o caráter composicional da língua, onde o conteúdo de todo o texto é a soma do conteúdo das partes. Depois de dividir o texto em partes léxicais, na maioria das vezes chamada de palavra ou termo, essas partes são ligadas umas às outras através da sintaxe, isto é, dicas do próprio texto de como elas se relacionam.

Por exemplo: "O cachorro mordeu João", indica que a palavra cachorro executou uma ação na palavra João. Se fosse "João mordeu o cachorro" indicaria contrariamente que o João executou uma ação no cachorro.

– Como o texto é representado?

As palavras são então agrupadas e classificadas segundo uma ontologia do conhecimento de forma que as seqüências que tiverem o mesmo significado apresentem a mesma representação. Na representação, são indicados alguns operadores lógicos de relacionamento como "contém", "pertence", "é igual a " etc.

Por exemplo: "O cachorro é um animal" indica que a palavra "cachorro" pertence à classe animal. Ou "Ver é o mesmo que Olhar" indica que a palavra "Ver" tem a mesma semântica que a palavra "Olhar".

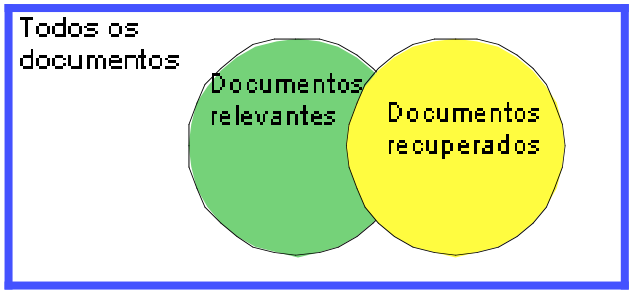
– Como compactar a informação?

No decorrer de um texto, são repetidas várias vezes a mesma informação, seja por repetição das palavras ou de sinônimos. O processo de agrupamento conceitual busca substituir todas as ocorrências que referenciam o mesmo conceito por um único identificador do conceito. Dessa forma a quantidade de caracteres iniciais é reduzida quando transformada em um modelo lógico de representação do conhecimento.

Por exemplo: em um texto podem ocorrer as palavras "PUC", "PUC-Rio", "Pontifícia Universidade Católica", "Faculdade" ou simplesmente "Universidade" e todas fazem referências ao mesmo objeto, podendo ser agrupadas em apenas uma entidade.

– Como medimos o sucesso de uma operação?

O critérios mais comumente utilizados na literatura são acurácia e abrangência, ou, precision e recall. Para calculá-los é necessário ter uma base de treinamento e teste apontando o certo e o errado, o relevante ou o irrelavante.



$$\text{Precisão} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos recuperados}}$$

$$\text{Eficiência (Recall)} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos relevantes}}$$

6. Quais os desafios?

– Grandes volumes de dados: Bases que ocupam em torno de 1 a 100GB.

– Alta dimensionalidade: Os problemas de mineração de textos normalmente manipulam um grande volume de palavras diferentes contidas em toda a coleção de textos.

– Super parametrização: Encaixar uma regra para uma dada frase, pode ser bastante simples, mas encontrar o número de regras que atendem a todas as possíveis construções da língua pode levar a uma quantidade não tratável de regras.

– Estruturas dinâmicas: Com o uso, novas regras são inventadas para nossa língua ao longo do tempo. Os sistemas de mineração de textos devem estar preparados para isso.

– Dados ruidosos: Freqüentemente são encontrados erros ortográficos nos textos analisados como palavras sem acento, ausência de espaço entre as palavras, tags html etc.

– Ambigüidade: Uma palavra pode atender a mais de um significado dependendo do contexto.

7. Técnicas utilizadas

A área de mineração de textos é uma área nova e multidisciplinar, que basicamente, além de algoritmos próprios, utiliza técnicas já conhecidas e consolidadas como:

– Indexação:

Serve para se fazer um rápida busca de documentos através de palavras-chave. Uma estrutura de dados de armazenamento inteligente proporciona aumento drástico de performance. Além de recuperar dados textuais, ela pode fazer cálculos com múltiplas palavras-chave de busca realizando uma ordenação segundo a avaliação de cada documento.

– Processamento de Linguagem Natural (PLN):

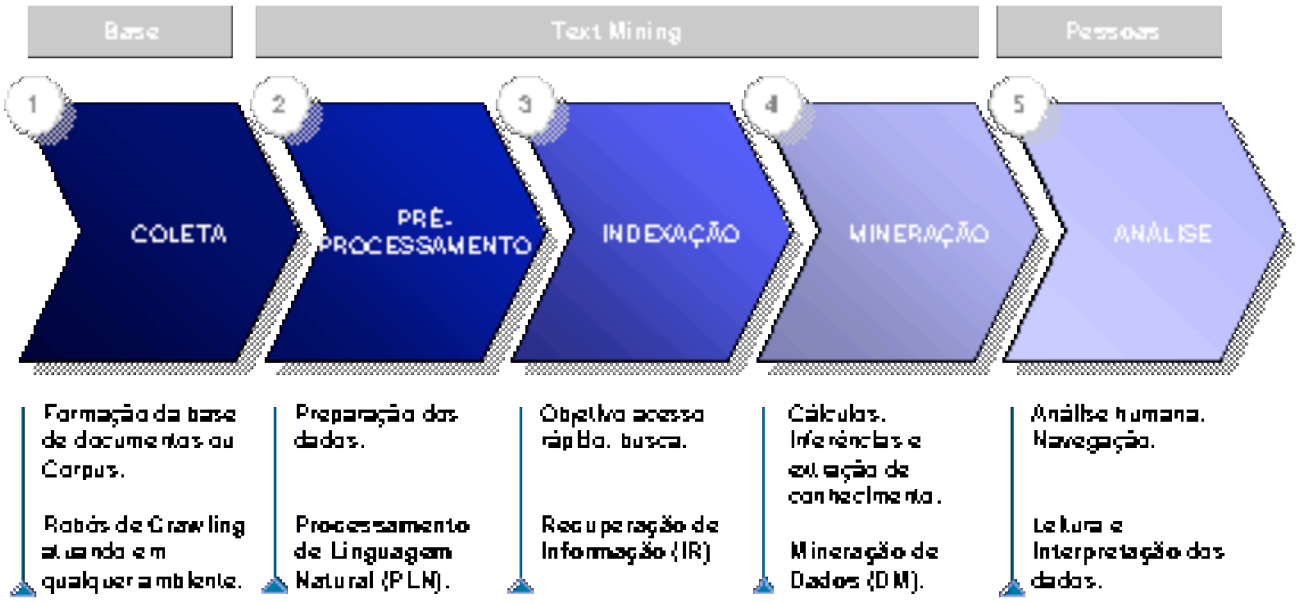
O processamento da linguagem natural (PLN) é outra técnica chave para mineração de textos. Utilizando conhecimentos da área de lingüística, o PLN permite aproveitar ao máximo o conteúdo do texto, extraindo entidades, seus relacionamentos, detectando sinônimos, corrigindo palavras escritas de forma errada e ainda desambiguizando-as. Participam normalmente na parte do pré-processamento dos dados, transformando-os em números.

– Mineração de dados

As técnicas inteligentes de Mineração de dados (“Data Mining”) são muito úteis para atuar em cima de um banco de dados organizado e pré-processado. Dessa maneira, é possível identificar os conhecimentos relevantes da base de dados textual. A técnicas mais utilizadas são Classificação, Clusterização e Otimização.

8. O Processo

O processo de Mineração de textos (“Text Mining”) como um todo, se constitui como mostra o diagrama a seguir:



Primeiramente é necessário compor uma base de textos a ser trabalhada, algumas vezes chamada de corpus. Normalmente o processo de mineração de textos começa a partir de uma base pronta, apesar de essa etapa ser uma das mais duras. Toda a base passa por um pré-processamento que vai estruturar o texto

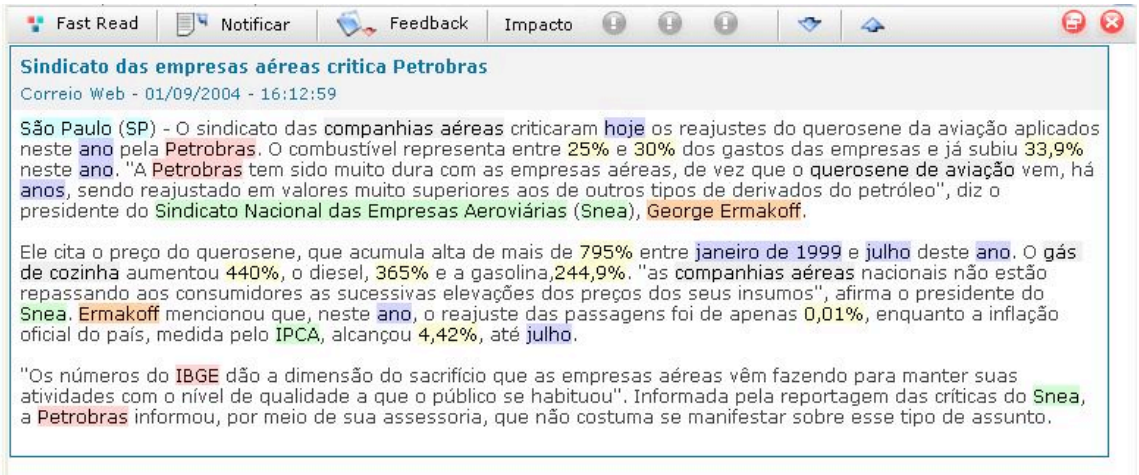
não-estruturado. Dependendo do volume de textos é necessário uma ferramenta de indexação e busca para acelerar o processamento dos textos. Sobre a estrutura resultante dessas etapas são aplicados algoritmos de mineração de dados para extrair os conhecimentos relevantes a serem analisados pelos usuários.

9. Exemplo de Pré-processamento

Utilizaremos um exemplo de texto como dado a seguir para demonstrar as funcionalidades e possíveis aplicações da tecnologia de mineração de textos.

> *São Paulo (SP) – O sindicato das companhias aéreas criticaram hoje os reajustes do querosene da aviação aplicados neste ano pela Petrobrás. O combustível representa ente 25% e 30% dos gastos das empresas e já subiu 33,9% neste ano. "A Petrobrás tem sido muito dura com as empresas aéreas, de vez que o querosene de aviação vem, há anos, sendo reajustado em valores muito superiores aos de outros tipos de derivados do petróleo", diz o presidente do Sindicato Nacional das Empresas Aeroviárias (Snea), George Ermakoff.*
>
> *Ele cita o preço do querosene, que acumula alta de mais de 795% entre janeiro de 1999 e julho deste ano. O gás de cozinha aumentou 440%, o diesel, 365% e a gasolina, 244,9%. "as companhias aéreas nacionais* não estão repassando aos consumidores as sucessivas elevações dos preços dos seus insumos", afirma o presidente do *Snea. Ermakoff mencionou que, neste ano, o reajuste das passagens foi de apenas 0,01%, enquanto a inflação oficial do país, medida pelo IPCA, alcançou 4,42%, até julho.*
>
> *"Os números do IBGE dão a dimensão do sacrifício que as empresas aéreas vêm fazendo para manter suas atividades com o nível de qualidade a que o público se habituou". Informada pela reportagem das críticas do Snea, a Petrobrás informou, por meio de sua assessoria, que não costuma se manifestar sobre esse tipo de assunto.*

– Reconhecimento de Entidades Nomeadas



As entidades pintadas de azul claro correspondem a lugares, as roxas são referências a tempo, as amarelas são números ou quantidades, as vermelhas são empresas e finalmente as em laranja são nomes de pessoas.

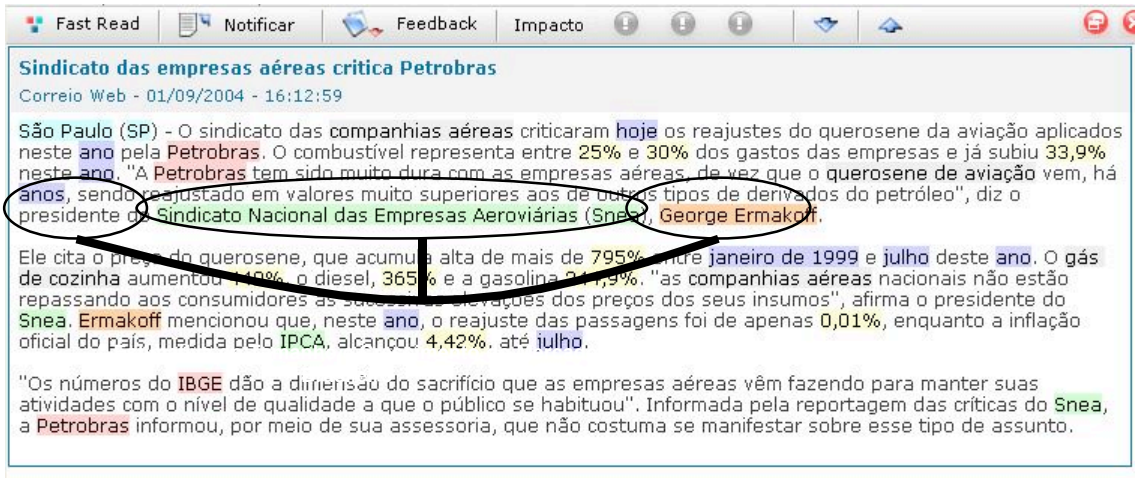
Nesse exemplo podemos destacar também a resolução de correferencias, onde a segunda ocorrência em laranja de "Ermakoff" está associada à primeira ocorrência "George Ermakoff", assim como o acrônimo Snea = "Sindicato Nacional das Empresas Aeroviárias". Vale ressaltar que os itens em verde não foram classificados.

– Extração de Entidades

O procedimento de extração de entidades visa organiza-las de modo a responder perguntas como Quem, Quando, Como, Onde. Para isso passam por uma etapa de normalização das entidades através de correferências, acrônimos e anáforas apresentadas no texto. Abaixo um exemplo de representação das entidades extraídas:

```
São Paulo = SP → Onde

hoje → Quando
neste ano → Quando

Petrobras → Organização
Sindicato Nacional das Empresas Aeroviárias = Snea → Organização
IBGE = Instituto Brasileiro de Geografia e Estatística → Organização

25%; 30%; 33,9%; 795%; 440%; 365%; 244,9%; 0,01%; 4,42% → Quantidade

George Ermakoff → Pessoa

IPCA = Índice de Preços ao Consumidor Amplo → Índice

companhias aéreas → Substantivo
querosene de aviação → Substantivo
reajuste → Substantivo
```

– Extração de Informação



A extração de informação consiste em associar logicamente as entidades previamente identificadas, reconhecidas e classificadas. No caso acima, um novo conhecimento foi adquirido, o de que "George Ermakoff" é o "presidente" do "Snea".

Todas essas informações podem ser armazenadas e representadas em formato XML, para que algoritmos inteligentes pré-concebidos de Data Mining possam ser aplicados para classificar os textos segundo as freqüências das entidades extraídas na fase de pré-processamento.

– Sumarização

A sumarização compreende dois processos: a seleção do conteúdo relevante de uma mensagem e sua organização coerente. Em ambos, três características principais devem ser observadas: a satisfação do objetivo comunicativo do discurso, a veiculação de sua proposição central e o inter-relacionamento coerente das unidades de conteúdo necessárias para satisfazer as restrições anteriores (ver Rino [15]).

Existem várias técnicas que podem ser utilizadas no processo de sumarização de textos. A seguir é

apresentado o resultado de uma utilizando frequencia das palavras como estimador de relevância.

O procedimento foi aplicado ao texto original:

"O sindicato das companhias aéreas criticaram hoje os reajustes do querosene da aviação aplicados neste ano pela Petrobrás. Ele cita o preço do querosene, que acumula alta de mais de 795% entre janeiro de 1999 e julho deste ano. As companhias nacionais não estão repassando aos consumidores as sucessivas elevações dos preços dos seus insumos"

10. Aplicações

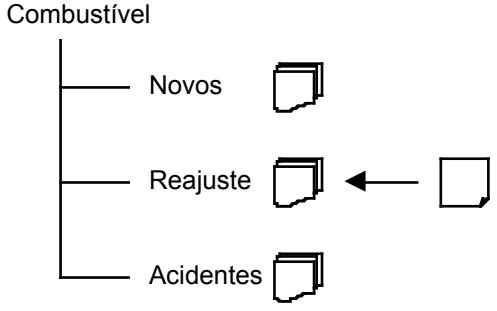
Para concretizar as idéias apresentadas, a seguir elas serão ilustradas por um exemplo concreto**.**

– Tipos de Classificação

	Palavras	Documentos
Aprendizado Supervisionado	Etiquetagem de ontologia, Desambigüização	Categorização, Filtragem, Detecção de tema
Aprendizado Não-supervisionado	Análise semântica Construção automática de taxonomia	Clustering de Documentos

– Classificação em Taxonomia

Efetua uma análise lingüística, onde os textos científicos, escritos em linguagem natural, fornecem a taxonomia conceitual das entidades significativas do discurso, com o auxílio da interpretação e do reconhecimento de palavras ou orações sinalizadoras da metalinguagem e de sinais lingüísticos.



O assunto principal do texto pode ser extraído e utilizado para organização de documentos em diretórios. Essa organização é muito útil para navegação.

– Extração de Regras

A partir de um conjunto de documentos, o sistema compara a freqüência das palavras de cada documento à freqüência de palavras em um conjunto de treinamento. O sistema guarda seu número de ocorrências, a partir da qual a palavra recebe um peso, relativo também ao número de ocorrências no corpus de treinamento. Palavras que ocorrem mais freqüentemente no documento do que no conjunto de treinamento recebem um peso maior.

Através desta técnica e da análise do banco de dados textual, é possível descobrir que 92% das vezes que aparece a instituição "Snea", aparece o nome "George Ermakoff". Sem que o usuário tenha lido nenhum texto da coleção, pode-se inferir que há uma forte relação entre ele e a instituição.

11. Softwares existentes na área

- Cortex Competitiva
Text Mining aplicado a Inteligência Competitiva.

- Text Analyst
Gera uma rede de semântica do texto baseada em um algoritmo de Hopfield.

- SAS Text Miner
Utiliza o famoso conjunto de ferramentas de Data Mining para Text Mining.

- Clementine
Utiliza as ferramentas do SPSS para Text Mining.

- Media Style
Apresenta soluções de extração de informação baseada em palavras-chave.

- Intext Mining
Text Mining Suite para Análise de Currículos

- WordStat
Os textos são categorizados automaticamente usando um dicionário de palavras.

Bibliografia

[1].ARANHA, C., Freitas, M. C., Dias, M. C., e Passos, E. (2004). "Um modelo de desambigüização de palavras e contextos". TIL 2004: Workshop de Tecnologia da Informação e da Linguagem Humana

[2].BAEZA-YATES, B. e Ribeiro Neto, B. (1999). Modern Information Retrieval. Addison Wesley

[3].BARCELOS, P. C. A. (2005). "Metodologia ou Tecnologia?". Anais do 12º Congresso Internacional ABED de Educação a Distância - "A Educação a Distância e a Integração das Américas", Florianópolis - SC

[4].BARRETO J. M.: Inteligência Artificial no Limiar do Século XXI, Capítulo 19, RôRôRô Edições, Florianópolis, 2001.

[5].BERRY, M. W. (2003). Survey of Text Mining: Clustering, Classification, and Retrieval. Springer

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

[7].DÖRRE, J., Gerstl, P., e Seiffert, R. (1999). "Text mining: Finding nuggets in mountains os textual data". San Diego, CA, pp.398-401.ACM

[8].FELDMAN, R. e Hirsh, H. (1997). "Exploiting Background information in Knowledge discovery from text". Journal of Intelligent Information System, v.9, n.1

[9].GOLDSCHMIDT R., & PASSOS, E.: Data Mining: um guia prático. Editora Campus, Rio de Janeiro: Elsevier, 2005

[10].GRUNE, D. e Jacobs, C. J. H. (1991). Parsing Techniques: A Practical Guide. Ellis Horwood Ltd

[11].HEARST, M. A. (1999). "Untangling Text Data Mining". Proceedings of the ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics.College Park: University of Maryland

[12].KANTROWITZ, M., Mohit, B., e Mittal, V. O. (2000). "Stemming and its effects on TFIDF ranking". SIGIR 2000: 357-359

[13].MANNING, C. e Schutze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press.

[14].PYLE, D. (1999). Data Preparation for Data Mining. San Francisco, CA: Morgan Kaufmann.

[15].RINO, L. H. M. (1996). "Sumarização automática de textos em português". Anais do II Encontro para o Processamento Computacional do Português Escrito e Falado, pp.109-119.Curitiba-PR.

[16].SAEED, J. L. (1997). Semantics. Oxford: Blackwell

[17].TAN, A.-H. (1999). "Text mining: The state of the art and the challenges". In Proceddings, PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, Beijing, pp.65-70

[18].WEISS, S. M., Indurkhya, N., Zhang, T., e Damerau, F. (2005). Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer Science+Business Media, Inc.

[19].ZANASI, A. (1997). Discovering Data Mining. Prentice Hall