Sistema para Manipulação de Textos baseado em Informação Semântica

Adriana C. Giusti Corrêa[1,2], Marina T. Pires Vieira[1], Marilde T. Prado Santos[1]

[1]Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
13565-905– São Carlos – SP – Brasil

[2]Instituto Municipal de Ensino Superior de Catanduva (FAFICA)
15800-020 – Catanduva – SP – Brasil

{adrianac,marina,marilde}@dc.ufscar.br

Abstract. *This paper presents a Text Manipulation System based on Semantic Information which utilizes text mining techniques to extract information which represent the semantic content of the text. Similarities between terms are extracted and stored in databases whose structure allows to carry out exact and fuzzy searches. The text Recovery differs from traditional methods for allowing the user to elaborate queries using a range of semantic information stored in the database and for using strategies for document ranking.*

Resumo. *Este artigo apresenta um Sistema para Manipulação de Textos baseado em Informação Semântica, no qual técnicas de mineração de textos são utilizadas para extrair informações que representam o conteúdo semântico destes. Dentre as informações extraídas destacam-se similaridades entre termos que são armazenadas em um banco de dados, cuja estrutura de classes oferece condições para a realização de buscas exatas e nebulosas. A recuperação dos textos diferencia-se dos métodos tradicionais por utilizar informações semânticas e ainda estratégias para a ordenação do resultado.*

1. Introdução

A mineração de textos surgiu da necessidade de análises automáticas em textos, visto que a sobrecarga de informações disponíveis dificultava sua análise manual, localização e acesso [Feldman & Dagan 1995].

Utilizando o processo de mineração em textos, foi desenvolvido um Sistema para Manipulação de Textos baseado em Informação Semântica que objetiva extrair informações semânticas de documentos em inglês e armazená-las em um banco de dados para posterior recuperação e ordenação dos resultados obtidos.

O artigo está organizado da seguinte forma: a seção 2 apresenta a arquitetura do sistema e os módulos componentes, descrevendo os recursos adotados em cada etapa, e ainda, o processo de recuperação dos documentos e ordenação do resultado. A seção 3 apresenta as conclusões finais do artigo.

2. Sistema para Manipulação de Textos Baseado em Informação Semântica

O Sistema para Manipulação de Textos tem por objetivo possibilitar a extração de informação semântica de documentos, assim como recuperar tais documentos através de buscas exatas e nebulosas com base na informação semântica. As buscas nebulosas são buscas que envolvem valores de similaridade mínima, composição de termos e relevância do termo para a busca. A ordenação dos documentos resultantes baseia-se em uma estratégia proposta por Corrêa [Corrêa 2003] que utiliza o grau de relevância do documento na expressão de busca

A arquitetura do sistema e os módulos componentes são apresentados na seção seguinte.

2.1. Arquitetura do Sistema

A arquitetura do sistema para manipulação de textos é apresentada na figura 1:

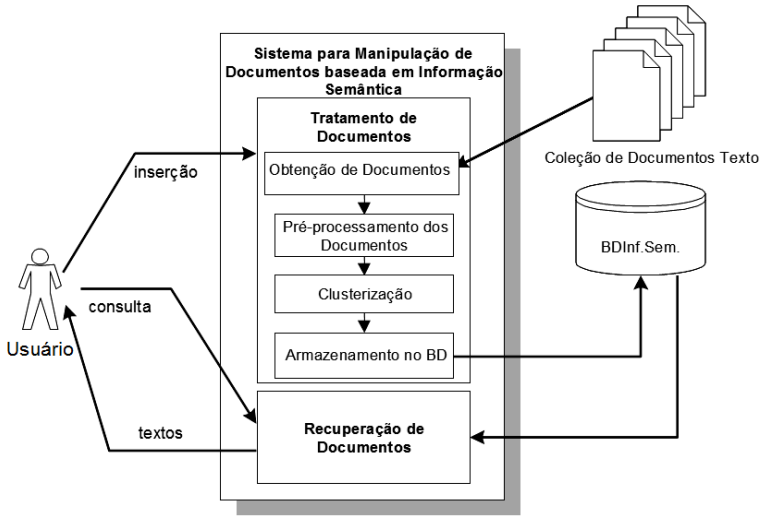

Figura 1: Arquitetura do Sistema para Manipulação de Textos

O sistema possui dois módulos principais que são apresentados nas seções seguintes.

2.2. Tratamento de Documentos

O módulo de Tratamento de Documentos contém as seguintes etapas:

1)Obtenção de Documentos: O usuário tem a opção de inserir documentos obtidos na Web por intermédio do aplicativo Wget [Wget 2001], o qual, neste trabalho, é utilizado para buscar arquivos com formato PDF e que contenham a palavra de busca definida pelo usuário, armazenando-os localmente.

2)Pré-processamento dos Documentos: O usuário seleciona os documentos obtidos na etapa anterior e submete-os ao pré-processamento. O objetivo desta etapa é obter um conjunto de palavras-chave que identificam o conteúdo de um documento, juntamente com os pesos e freqüências associados a elas. Para tanto, as seguintes tarefas são executadas:

a. Transformação do documento para o formato texto**;**

b. Limpeza e Padronização do Texto: Dígitos são retirados e é feita a conversão de todos os caracteres do texto para minúsculo;

c. Remoção de *stop-words*: Cada palavra do documento é comparada com as palavras contidas em uma *stoplist,* que neste trabalho foi obtida do sistema KEA 2.0, desenvolvido por Frank [Frank et. al. 2000]. As palavras que são encontradas na *stoplist* devem ser retiradas do documento, pois não são representativas de seu conteúdo;

d. *Stemming*: O algoritmo de *stemming* faz com que as palavras que restaram do processo anterior sejam transformadas em sua forma raiz, diminuindo a quantidade de palavras diferentes contidas em um documento. O algoritmo utilizado é o Porter Stemmer [Porter 1980].

e. Determinação de pesos: Para cada palavra restante deve ser associado um peso indicando a relevância da palavra dentro do texto e na coleção. Para calcular o peso são utilizadas as fórmulas propostas por [Salton & McGill 1983]:

$$WEIGHT_{ik} = FREQ_{ik} * IDOCFREQ_k \text{ e } IDOCFREQ_k = \log_2 \frac{n}{DOCFREQ_k} + 1$$

onde $FREQ_{ik}$ é a freqüência de ocorrência de cada termo k em cada documento i e $IDOCFREQ_k$ corresponde à freqüência inversa do termo k.

f. Seleção de palavras-chave: as palavras com fatores de peso mais altos são atribuídos como palavras-chave aos textos.

3)Clusterização: A clusterização é uma das possíveis tarefas utilizadas na mineração em textos e caracteriza-se por agrupar em classes os termos que possuem um certo grau de similaridade. Para isso, são calculados os valores de similaridade entre todos os pares de termos através da fórmula, sugerida por Salton et. al. [Salton &McGill 1983]:

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

onde w_{ik} indica o peso do $TERMO_k$ no documento i e assumindo n documentos na coleção.

4)Armazenamento no Banco de Dados: Toda a informação semântica (palavras-chave, composição de termos, similaridade entre termos e classes de termos) é armazenada e utilizada em buscas exatas ou nebulosas. O Sistema Gerenciador de Banco de Dados utilizado é o Caché [Caché 2002], que suporta objetos, facilitando a manipulação de estruturas complexas. A estrutura de classes que permite armazenar a informação semântica pode ser vista na figura 2. Essa estrutura de classes foi desenvolvida por Vieira et. al. [Vieira et. al. 2002] e foi estendida para atender ao tratamento de informações semânticas de documentos texto, segundo a abordagem adotada neste trabalho [Corrêa 2003].

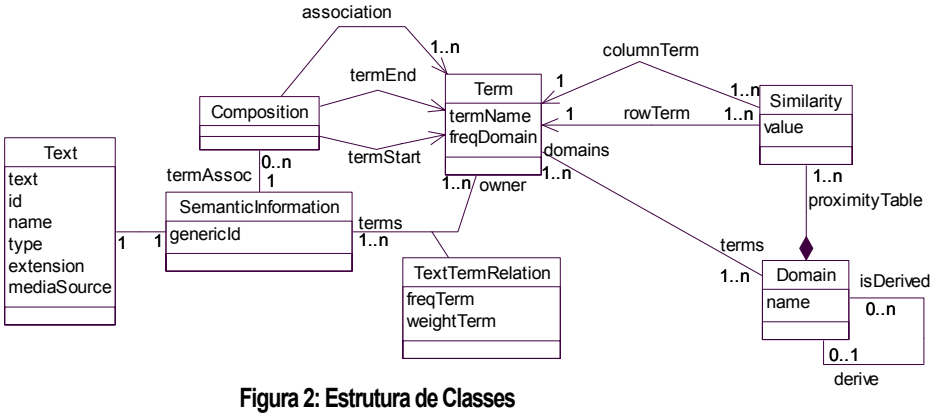

Figura 2: Estrutura de Classes

A classe *Text* permite armazenar informações de textos, tais como, o nome, tipo, extensão, a localização do texto, e o seu conteúdo. Cada texto possui uma representação de Informação Semântica (*SemanticInformation*). A classe *SemanticInformation* está relacionada com uma ou mais palavras-chave extraídas do texto. Tais palavras-chave, tratados aqui como termos, estão representadas na classe *Term*. As informações sobre a freqüência e o peso do termo são armazenadas em *TextTermRelation*. Os termos pertencem a um domínio e, para tanto, a classe *Domain* armazena o relacionamento entre os termos, expresso através de valores de similaridade (*value*, da classe *Similarity*). A classe *Domain* permite representar domínios e subdomínios de termos obtidos pela similaridade entre os pares de termos. A classe *Composition* adiciona características no tratamento dos textos, pois é possível formar composições de termos, onde existem o termo de início (*termStart*) e o termo final da composição (*termEnd*). Alguns exemplos de informação semântica, juntamente com a recuperação de documentos são contemplados na seção seguinte.

2.3. Recuperação de Documentos

A recuperação de documentos envolve características especiais por permitir a elaboração de buscas exatas ou nebulosas. As buscas nebulosas envolvem termos similares ao termo original definido pelo usuário e ainda valores de relevância que são utilizados no momento da ordenação do resultado obtido. Um conjunto de fórmulas foi definido para a ordenação dos documentos pertencentes ao conjunto resposta; maiores detalhes sobre a estratégia utilizada podem ser encontrados em [Corrêa 2003].

Dentre as possíveis expressões de busca, é apresentada a seguir uma expressão de busca nebulosa, envolvendo composição de termos, similaridade mínima para um determinado termo e valores de relevância.

Hierarchy [100%,100%] WITH classification [50%,100%]

Os valores entre colchetes representam respectivamente a similaridade mínima e a relevância do termo, ambas especificadas pelo usuário. Isto representa que qualquer termo que seja, no mínimo, 50% similar a "classification" será considerado na busca, enquanto que a relevância é utilizada para ordenar os documentos recuperados. A composição de termos pode ser observada pelo termo de associação (WITH) que adiciona uma semântica diferenciada à expressão de busca, relacionando o termo de início (hierarchy) e o termo final (classification) da composição. O resultado para a expressão de busca pode ser visto a seguir:

Documento	Weight hierarchy	Relevance hierarch	Termo Similar / weight	GR_{Di}
1) Vscluster.pdf	0.73	1	-	0.58
2) Clustering.pdf	0.49	1	-	0.52
3) Beil02frequent.pdf	0.53	1	"agglom"/0.79	0.66

Os documentos estão ordenados de acordo com o grau de relevância calculado a partir da fórmula:

$$GR_{Di,Gj} = \frac{GR_{t_k,Di} + GR_{t_{k+1},Di}}{relevance_{t_k} + relevance_{t_{k+1}}}$$

obtida através do peso do termo para o documento e relevância especificada. O terceiro documento apresentado aparece por último na classificação pois atende à busca nebulosa, contendo um termo similar ao termo original.

3. Conclusões

Este artigo apresentou um Sistema de Manipulação de Textos baseado em Informação Semântica, propondo uma forma de auxiliar a extração de informações relevantes de textos, armazenando-as em um banco de dados. A recuperação dos documentos caracteriza-se por realizar buscas exatas ou nebulosas, envolvendo a informação semântica armazenada e ordenando o resultado de acordo com o grau de relevância do documento para a expressão de busca.

Técnicas de mineração de textos foram utilizadas para extração de palavras-chave e determinação de similaridades entre pares de termos, possibilitando a criação de classes de termos similares.

Os autores pretendem comparar resultados da recuperação em relação à abordagem tradicional de Recuperação de Informação Textual, relatando as vantagens das técnicas utilizadas no sistema em questão.

6. Referências

Caché, Intersystems (2002). Banco de Dados Pós-Relacional, versão 4.1.6 – Disponível em: http://www.intersystems.com.br. Julho, 2002.

Corrêa, A. C. G. (2003) "Recuperação de Documentos baseada em Informação Semântica no Ambiente AMMO". Dissertação de Mestrado em Ciência da Computação. Universidade Federal de São Carlos.

Feldman, R.; Dagan, I. (1995) "Knowledge Discovery in Textual Databases (KDT)". In First International Conference on Knowledge Discovery (KDD´95), Montreal.

Frank, E., Witten I.H., Paynter G.W. (2000) "KEA: Practical automatic keyphrase extraction." Department of Computer Science, The University of Waikato.

Porter, M. F. (1980) "An algorithm for suffix stripping". Readings in Information Retrieval, San Francisco: Morgan Kaufmann.

Salton, G.; McGill, M. (1983) "Introduction to Modern Information Retrieval". New York: McGraw-Hill.

Vieira, M.T.P. et. al. (2002) "Content-Based Fuzzy Search in a Multimedia Web Database". In Intelligent Exploration of the Web, "Studies in Fuzziness and Soft Computing" series, Springer-Verlag.

Wget Copyright (2001) Free Software Foundation. http://www.gnu.org/software/wget/wget.html. Dezembro.