Informações sintáticas na Mineração de Textos

Cassiana da Silva, Cláudia Perez, Fernando Osório,
Renata Vieira, Rodrigo Goulart

Programa Interdisciplinar de Pós-Graduação em Computação Aplicada
Universidade do Vale do Rio dos Sinos
{cassiana, claudiap, osorio, renata, rodrigo}@exatas.unisinos.br

Abstract. *In this paper we show two experiments that use syntactic information in text mining tasks. The PALAVRAS parser provides the syntactic information used in our experiments.*

Resumo. *Este artigo apresenta dois experimentos que usam informação sintática em tarefas de mineração de textos. O parser PALAVRAS fornece a informação sintática utilizada nos experimentos.*

1. Introdução

Este trabalho descreve dois experimentos de Mineração de Textos baseados em informação morfossintática da língua portuguesa do Brasil, envolvendo as tarefas Extração de Informação e Categorização de Textos. A Extração de Informação (EI) é a tarefa de extrair informação relevante de um texto em linguagem natural, e apresentá-las em uma estrutura formal. Estas informações podem ser utilizadas posteriormente na execução de uma tarefa particular, por um sistema computacional. A Categorização é uma técnica empregada para identificar a classe ou categoria a que um determinado documento pertence, utilizando como base o seu conteúdo. Para tanto, as classes devem ter sido previamente modeladas ou descritas através de suas características, atributos ou fórmula matemática.

2. Informação sintática

A análise sintática utilizada nos experimentos é fornecida pelo parser PALAVRAS desenvolvido para o português por Eckhard Bick (Bick 2000) no projeto VISL (*Visual Interactive Syntax Learning*)[1]. Em (Gasperin 2003) a geração de elementos XML correspondentes à análise do PALAVRAS, utilizada em nossos experimentos, é apresentada. Três arquivos em formato XML são gerados: O primeiro contém as palavras do corpus, o segundo as informações morfo-sintáticas das palavras, e o terceiro contém informação sobre a estrutura sintática das sentenças (a Figura 2.1 apresenta o arquivo de palavras e suas estruturas em XML, onde s = sujeito, v = verbo, od = objeto direto, h = núcleo do sintagma). Com a informação codificada em XML, podemos aplicar folhas de estilo XSL[2] para a extração das estruturas sintáticas a serem utilizadas nas aplicações. A seguir são apresentados dois experimentos de mineração de textos, baseados nas informações obtidas pelo processo descrito acima.

[1] http://visl.sdu.dk/visl/pt/parsing/automatic/
[2] *eXtensible Stylesheet Language* http://www.w3.org/Style/XSL/

```
<word id="word_141">Oliveira</word>
<word id="word_142">deixou</word>
<word id="word_143">o</word>
<word id="word_144">cargo</word>
<word id="word_145">anteontem</word>
<word id="word_146">para</word>
<word id="word_147">concorrer</word>
<word id="word_148">a</word>
<word id="word_149">deputado</word>
<word id="word_150">federal</word>
<word id="word_151">por</word>
<word id="word_152">o</word>
<word id="word_153">PSDB</word>
<word id="word_154">.</word>
```

```
<chunk id="chunk_78" ext="s" span="word_141">
</chunk>
<chunk id="chunk_79" ext="v" span="word_142">
</chunk>
<chunk id="chunk_80" ext="od" span="word_143..word_144">
<chunk id="chunk_81" ext="h" span="word_144">
</chunk>
</chunk>
```

Figura 2.1 Elementos XML gerados

3. Extração de Informação

Como um primeiro estudo, desenvolveu-se um trabalho sobre a tarefa de aquisição conhecimento, através de um processamento semi-automático (Pérez 2003). O conhecimento extraído é representado por Mapas Conceituais, que são recursos para a representação de conhecimento, constituídos numa rede de nós, representando os conceitos, conectados por arcos com rótulos, representando as relações entre pares de nós (Araújo 2002, Cañas 2000, Ford 1991). A relação entre dois conceitos é dada por palavras que evidenciam o por quê de um relacionamento entre conceitos e seus relacionamentos são representados por triplas *conceito – relação – conceito*, que em textos da língua natural tendem a aparecer como *sujeito – verbo – objeto*(<s-v-o>). Nesse experimento avaliamos a extração automática das triplas <s-v-o> de textos para a identificação das relações (verbos) entre os conceitos (sujeito e objeto).

A Figura 3.1 ilustra parte do mapa conceitual obtido automaticamente. Para a geração de grafos foi utilizada a ferramenta Cmap Tools, desenvolvida pelo IHMC- University of West Florida (IHMConcept 2003). Nesse primeiro experimento, são extraídos apenas os núcleos dos sujeitos, verbos e objetos de textos jornalísticos. Mas em vários casos extraídos, como por exemplo, pronomes que exercem o papel de sujeito ou objeto, verbos com complementos proposicionais (dizer, afirmar), nomes compostos, as relações formadas apenas com os núcleos são insuficientes. Além disso, o corpus, constituído de artigos do jornal Folha de São Paulo, possui um estilo de escrita e conteúdo, relacionado a acontecimentos, que dificulta a aquisição de conhecimento próprios do estilo dos mapas conceituais.

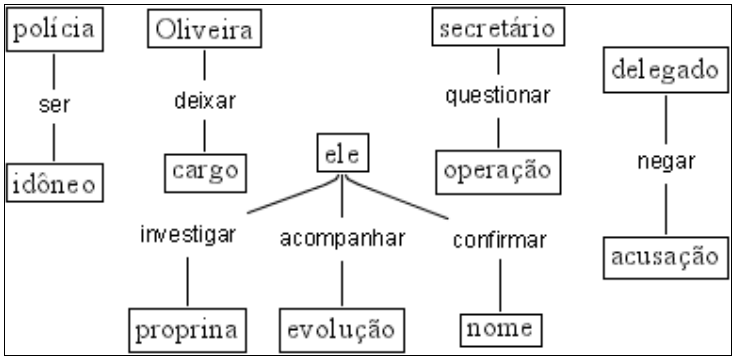

Figura 3.1 - Mapa Conceitual

4. Categorização de textos

Nosso segundo experimento é relacionado à avaliação do uso de informações sintáticas na seleção de atributos para categorização de textos. Os experimentos foram realizados usando um corpus de texto puro e um corpus contendo anotações morfo-sintáticas.

Para a realização dos experimentos com o corpus de textos puros, os textos foram primeiramente submetidos ao pré-processamento tradicional de remoção de stopwords e lematização. O corpus contempla 117 documentos, divididos em um conjunto de treinamento (2/3 dos exemplos) e um conjunto de teste (1/3 dos exemplos). Os experimentos realizados com o corpus contendo informações morfo-sintáticas, o pré-processamento foi baseado em informações lingüísticas e com isso, optou-se pela seleção dos substantivos dos textos como atributos para as categorias. Ambos experimentos utilizam a seleção dos atributos baseada na freqüência mínima dos termos (5 e 7). O número de palavras relevantes selecionadas (atributos) do corpus foi de 180, 148 e 148, 64 respectivamente.

Os experimentos de categorização foram feitos com as árvores de decisão, utilizando-se a ferramenta C4.5 (Quinlan 1993) e com redes neurais, através do Neusim (Osório 1999). A RNA utilizada foi a MLP, sendo aplicados os algoritmos de aprendizado *Backpropagation* e *Cascade-Correlation.* A Tabela 4.1 apresenta o erro de classificação dos exemplos utilizando as técnicas simbólicas e conexionistas, utilizando codificação por freqüência.

Tabela 4.1 Erro de Classificação dos exemplos do Corpus Original e do Corpus Substantivos

Tipo deCorpus	**Original**		**Substantivos**	
Número de Atributos	**180**	**148**	**148**	**64**
RNA *Backpropagation*	15%	5%	5%	15%
RNA *Cascade Correlation*	7.5%	15%	10%	15%
Árvores de Decisão	17.1%	15%	25%	27,5%

Com base nos resultados apresentados na Tabela 4.1, podemos observar que a melhor generalização dos exemplos no corpus original foi obtida utilizando as RNAs com o algoritmo *Backprop* (5%), com um número de 148 atributos relevantes. No

corpus substantivos conclui-se que é possível encontrar uma generalização tão boa como os métodos tradicionais de pré-processamento

O experimento realizado com o corpus inverso ao substantivo visava provar que a seleção de todas as palavras que não eram substantivos do corpus apresentaria uma generalização ruim. Com base nos resultados obtidos, observa-se que a queda esperada no resultado da classificação não ocorreu. Logo, uma análise nos atributos selecionados foi realizada com o intuito de descobrir informações que identificassem o resultado obtido. Nessa análise verificou que além de stopwords e verbos existiam nomes próprios e adjetivos (tais como, Palmeiras, Botafogo, Fernandinho, entre outros).

Este segundo estudo propôs uma avaliação na seleção dos atributos no processo de categorização de documentos. Duas abordagens foram adotadas, uma usando atributos extraídos de um corpus não anotado em relação a atributos extraídos de um corpus com anotação morfo-sintática. Os experimentos realizados com o corpus com anotação morfo-sintática apresentaram um ganho na seleção dos atributos, concluindo que a seleção dos substantivos é tão satisfatória quanto o processo tradicional de remoção de stopwords e lematização na categorização de documentos.

5. Conclusões

Esse artigo apresentou dois experimentos de mineração de textos com base em informações sintáticas. Os resultados preliminares apresentados indicam que a identificação de determinadas estruturas podem auxiliar tanto no processo de extração de informação como na categorização de documentos.

6. Blibiografia

Araújo, A., Menezes, C., and Cury, D. (2002) “Um Ambiente Integrado para Apoiar a Avaliação da Aprendizagem Baseado em Mapas Conceituais”. Anais do XII Simpósio Brasileiro de Informática na Educação (SBIE 2002), p. 49-58.

Bick, E. (2000) “The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework”. In: PH. D. thesis, Arhus University, 2000.

Cañas, A., Ford, K., Coffey, J., et al. (2000) “Herramientas para Construir y Compartir Modelos de Conocimiento Basados en Mapas Conceptuales”. Revista: Informática Educativa, Vol. 13, No. 2, p. 145-158.

Ford, K., Cañas, A., Jones, J., Stahl, H., Novak, J., et al. (1991) “ICONKAT: an Integrated Constructive Knowledge Acquisition tool”. Academic Press Limited, 1991.

Gasperin, C., Vieira, R., Goulart, R. and Quaresma, P. (2003) “Extracting XML Syntactic Chunks from Portuguese Corpora”. In: Proceedings of the Workshop TALN 2003 Natural Language Processing of Minority Languages and Small Languages, France, June, 2003.

IHMConcept Map Software a knowledge construction toolkit (2003). Software disponível em http://cmap.coginst.uwf.edu/, Maio.

Osório, F. and Amy, B. (1999) “INSS: A hybrid system for constructive machine learning”. Neurocomputing 28: 191-205, 1999. Software disponível em http://www.inf.unisinos.br/~osorio/INSS.

Pérez, C., Gasperin, C. and Vieira, R. (2003) Extração Semi-Automática de Conhecimento. Anais do Encontro Nacional de Inteligência Artificial (ENIA 2003), Campinas, SP, 4 - 8 Agosto, 2003.

Quinlan, J. R. (1993) “C 4.5 : Programs for Machine Learning”. San Mateo: Morgan Kufmann Publishers, 1993.