XII Salão de Iniciação Científica PUCRS

Um Estudo sobre Categorização de Textos usando a WordNet

Renato Falsarella Malvezzi, Gustavo Barcellos Miranda, Ricardo Manhago Serro, Silvia Maria Wanderley Moraes (orientador)

Faculdade de Informática, PUCRS

Resumo

Este artigo descreve um estudo sobre categorização de textos enriquecidos com sinônimos da WordNet, realizado com textos do *corpus* Reuters-21578. Foram implementadas duas abordagens para indexação dos documentos. Na primeira, estes foram representados usando-se apenas os termos mais freqüentes. Na segunda, incluímos também os sinônimos destes termos. Utilizamos dois classificadores disponíveis na ferramenta Weka: o KStar e a rede neural Multi-Layer Perceptron, analisando e comparando os resultados com um trabalho relacionado

Introdução

Segundo Sebastiani em (SEBASTIANI, 2005), categorizar textos consiste em organizar um conjunto de documentos em categorias previamente conhecidas. A área é de constante pesquisa em busca soluções que ofereçam maior agilidade e precisão (AGHDAM *et al*, 2008), dado o grande volume de documentos eletrônicos que surgem diariamente. A fim de melhorar os resultados, pesquisadores têm incluído no processo de categorização recursos lexicais e estruturas conceituais, como a WordNet[2]. O trabalho de Elberrichil et al em (ELBERRICHIL et al, 2006) é um exemplo disso. Os autores usam a WordNet para buscar sinônimos e hiperônimos das palavras do texto durante a etapa de indexação dos documentos, capturando as relações entre elas. Aplicam ainda a medida X^2 para estabelecer o grau de associação entre os termos e as categorias dos documentos. A medida X^2 ajuda a reduzir a dimensionalidade de representação dos documentos, pois termos com baixos X^2 são descartados. Os pesos dos termos são calculados usando a medida TFIDF. Os autores usaram em seus experimentos o *corpus* Reuters-21578 e dados do 20 Newsgroups, e a medida co-seno para determinar a similaridade entre os

[2] Disponível em http://wordnet.princeton.edu/

documentos e as categorias. Com o uso da WordNet, eles conseguiram aumentar a macro-média F1 de 0,649 para 0,714 no caso da Reuters-21578 e de 0,667 para 0,719 no caso de 20 Newsgroups.

Metodologia

A fim de que pudéssemos comparar nossos resultados com os de Elberrichil *et al* em (ELBERRICHIL et al, 2006), usamos 9.545 documentos do corpus Reuters-21578[3], considerando as mesmas categorias que os autores. São elas: "earn"(3.775 documentos), "acq"(2.210), "money-fx"(682), "grain"(573), "crude"(564), "trade"(515), "interest"(422), "wheat"(287), "ship"(294) e "corn"(223). Para cada categoria, 70% dos textos foram escolhidos para o conjunto de treino, ou seja, 6.686 documentos. O restante, 2.859, formaram o conjunto de teste. Os textos foram lematizados[4] usando a ferramenta TreeTagger[5]. Foram eliminadas as chamadas palavras negativas, também conhecidas por *stopwords*, que correspondem a artigos, preposições, conjunções, etc. Foram excluídos também caracteres especiais, números e o verbo *to be*. A partir do conjunto de treino foram identificadas as 50 palavras mais freqüentes em cada categoria. Com a união dessas palavras formou a *bag-of-words* usadas pelos classificadores. Para os documentos, foi utilizada uma representação binária, indicando apenas ausência (0) ou presença (1) do termo da *bag-of-words*. Essa foi a primeira abordagem testada. A segunda abordagem consistiu em enriquecer a *bag-of-word* com sinônimos obtidos a partir da WordNet. Foram considerados apenas os sinônimos contidos no primeiro *synset*. A idéia era usar apenas o significado mais usual das palavras. Após isso, os documentos foram submetidos aos classificadores K-Star e MultiLayer Perceptron (MLP) disponíveis na ferramenta Weka.

Cabe ressaltar ainda que a etapa de preparação dos documentos foi implementada tanto em Java quanto em Phyton. Observamos que a linguagem Phyton é mais adequada para esse tipo de processamento. A velocidade de desenvolvimento e de processamento foram maiores com o uso de Phyton.

Forma de Avaliação e Resultados

Para comparar os resultados das diversas soluções usamos as medidas conhecidas *Precision (Pr)*, *Recall (Re)* e *F-Measure (F1).* A Tabela I descreve os resultados que obtivemos

[3] Disponível em http://www.daviddlewis.com/resources/testcollections/reuters21578/
[4] A lematização é o processo de transformar uma palavra em sua forma canônica, ou seja, os verbos são colocados no infinitivo e os substantivos no masculino-singular.
[5] Disponível em http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

por categoria considerando os dois classificadores K-Star e MLP, com e sem o uso da WordNet, com implementação em Phyton.

Tabela I -Resultados de categorização de nosso estudo.

Categoria	**KStar sem WordNet**			**KStar com WordNet**			**MLP sem WordNet**			**MLP com WordNet**		
	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
acq	0,849	0,926	0,886	0,849	0,926	0,886	0,950	0,940	0,945	0,961	0,935	0,948
corn	0,364	0,060	0,103	0,364	0,060	0,103	0,421	0,119	0,186	0,50	0,015	0,029
crude	0,776	0,657	0,712	0,776	0,657	0,712	0,754	0,799	0,776	0,785	0,734	0,758
earn	0,947	0,979	0,963	0,947	0,979	0,963	0,971	0,981	0,976	0,97	0,984	0,977
grain	0,475	0,703	0,567	0,475	0,703	0,567	0,475	0,779	0,590	0,48	0,785	0,596
interest	0,579	0,520	0,548	0,579	0,520	0,548	0,703	0,504	0,587	0,727	0,441	0,549
money-fx	0,657	0,634	0,645	0,657	0,634	0,645	0,668	0,746	0,705	0,649	0,839	0,732
ship	0,595	0,534	0,563	0,595	0,534	0,563	0,580	0,659	0,617	0,564	0,705	0,626
trade	0,700	0,773	0,735	0,700	0,773	0,735	0,750	0,760	0,755	0,718	0,825	0,767
wheat	0	0	0	0	0	0	0,375	0,035	0,064	0	0	0
médias	0,783	0,810	0,792	0,783	0,810	0,792	0,835	0,840	**0,828**	0,828	0,842	0,824

Analisando a Tabela I percebemos que os melhores resultados foram obtidos com o classificador MLP. Embora, o F1 para categorização usando MLP sem WordNet, na média, tenha sido um pouco mais alto, podemos observar que o uso da WordNet trouxe benefícios ao processo. Para categorias como "acq" e "com" houve aumento na precisão. Já para categorias, como "ship" e "trade", o ganho foi na abrangência (Recall). Comparando ainda nossos resultados com os de ELBERRICHIL *et al*, observamos uma melhora significativa. Enquanto que em média os autores conseguiram F1 de 0,667 para categorização sem o uso da WordNet e F1 de 0,719 usando WordNet, em nosso estudo atingimos índices maiores. Conseguimos 0,828 e 0,824, respectivamente, para categorização usando MLP sem e com WordNet.

Conclusão

Apesar dos resultados, em média, para o índice F1 terem sido um pouco maiores para a categorização sem a WordNet, consideramos os resultados promissores. Cabendo, portanto, a continuidade da pesquisa com a WordNet, utilizando-se, por exemplo, outros classificadores bem como outras abordagens para construção *da bag-of-words*.

Referências

AGHDAM, M.H ; GHASEM-AGHAEE, N.; EHSAN BASIRI, M., **Application of ant colony optimization for feature selection in text categorization**. In Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). Hong Kong, Japão, 2008.

ELBERRICHIL, Z.; RAHMOUN, A., e BENTAALAH M. **Using WordNet for Text Categorization** - 1EEDIS Laboratory, Department of Computer Science, University Djilali Liabès, Algeria King Faisal University, Saudi Arabia, 2006.

SEBASTIANI, F. **Text categorization**. In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK, pp. 109-129, 2005.