西安电子科技大学学报 2013, 40(2) 89-97+129 DOI: 10.3969/j.issn.1001-2400.2013.02.015 ISSN: 1001-2400 CN: 61-1076/TN

研究论文

面向文本分类的中文文本语义表示方法

宋胜利;王少龙;陈平

(西安电子科技大学 软件工程研究所，陕西 西安 710071)

摘要：

为了解决词频统计文本表示方法中词语间语义信息缺失的问题，在考虑文本中词语上下文语境和语义背景信息的基础上，提出了一种新的中文文本表示模型——文本语义图．该方法利用维基百科作为知识背景计算文本中实意特征词语的语义关联，将具有较强语义关系的词语合并成词包作为图的节点，节点权值用词包所包含词语的数目及词频计算; 不同词包中词语间的上下文关系作为图的有向边，有向边权值用其邻接节点的最大权值表示．该模型在较大程度地保留文本中词语上下文信息的同时强化了词语间语义内涵．通过中文文本分类实验，文本语义图分类方法相对于支持向量机分类效率提升了7.8%，同时错误率减少了1/3，且表现出更好的稳定性．实验结果表明在文本分类应用中，文本语义图模型能够有效地表示文本内容．

关键词： 分类 知识表示 相似度 文本语义图

Chinese text semantic representation for text classification

SONG Shengli;WANG Shaolong;CHEN Ping

(Research Inst. of Software Engineering, Xidian Univ., Xi'an 710071， China)

Abstract:

Text representation based on word frequency statistics is often unsatisfactory because it ignores the semantic relationships between words, and considers them as independent features. In this paper, a new Chinese text semantic representation model is proposed by considering contextual semantic and background information on the words in the text. The method captures the semantic relationships between words using Wikipedia as a knowledge base. Words with strong semantic relationships are combined into a word-package as indicated by a graph node, which is weighted with the sum of the number and frequency of the words it contains. The contextual relationship between words in different word-packages is stated by a directed edge, which is weighted with the maximum weight of its adjacent nodes. The model retains the contextual information on each word with a large extent. Meanwhile, the semantic meaning between words is strengthened. Experimental results of Chinese text classification show that the proposed model can express the content of a text accurately and improve the performance of text classification. Compared to Support Vector Machines, Text Semantic Graph-based Classification can improve the efficiency by 7.8%, reduce the error rate by 1/3, and show more stability.

Keywords: classification knowledge representation similarity text semantic graph

收稿日期 2011-11-11 修回日期 网络版发布日期

DOI: 10.3969/j.issn.1001-2400.2013.02.015

基金项目:

国家自然科学基金资助项目(JJ0500092301)；中央高校基本科研业务费资助项目(K50510230003)

通讯作者: 宋胜利

作者简介: 宋胜利(1981-)，男，讲师，博士，E-mail: shlsong@xidian.edu.cn.

作者Email: shlsong@xidian.edu.cn

参考文献：

［1］ Li Yuhua, Mclean D, Bandar Z A, et al. Sentence Similarity Based on Semantic Nets and Corpus Statistics ［J］. IEEE Trans on Knowledge and Data Engineering, 2006,18(8): 1138-1150.

［2］ Schenker A, Last M, Bunke H, et al. Classification of Web Documents Using a Graph Model ［C］//Proc of the 7th International Conference on Document Analysis and Recognition. Washington: IEEE Computer Society, 2003: 240-244.

［3］ 吴江宁, 刘巧凤. 基于图结构的中文文本表示方法研究 ［J］. 情报学报, 2010, 29(4): 618-624.

Wu Jiangning, Liu Qiaofeng. Research on Graph Structure Based Method for Chinese Text Representation ［J］. Journal of The China Society for Scientific and Technical Information, 2010, 29(4): 618-624.

［4］ Manuel M G, Aurelio L L, Alexander G. Information Retrieval with Conceptual Graph Matching［C］//Proc of the 11th International Conference on Database and Expert Systems Applications. London: Springer-Verlag, 2000: 312-321.

［5］ Bhoopesh C, Pushpak B. Text Clustering Using Semantics ［C］//Proc of the 11th International Conference on World Wide Web. New York: ACM Press, 2002: 79.

［6］ Svetlana H. Construction of Conceptual Graph Representation of Texts ［C］//Proc of Student Research Workshop at HLT-NAACL 2004. Stroudsburg: Association for Computational Linguistics, 2004: 49-54.

［7］ Song W, Park S C. A Novel Document Clustering Model Based on Latent Semantic Analysis ［C］//Proc of the 3rd International Conference on Semantics, Knowledge and Grid. Washington: IEEE Computer Society, 2007: 539-542.

［8］ Lee C S, Kao Y F, Kuo Y H, et al. Automated Ontology Construction for Unstructured Text Documents［J］. Data ＆ Knowledge Engineering, 2007, 60(3): 547-566.

［9］ Stavrianou A, Andritsos P, Nicoloyannis N. Overview and Semantic Issues of Text Mining ［J］. ACM SIGMOD Record, 2007, 36(3): 23-34.

［10］ Jin W, Srihari R K. Graph-based Text Representation and Knowledge Discovery ［C］//Proc of the 2007 ACM Symposium on Applied Computing. New York: ACM Press, 2007: 807-811.

［11］ Chang M W, Ratinov L, Roth D, et al. Importance of Semantic Representation: Dataless Classification ［C］//Proc of the 23rd AAAI Conference on Artificial Intelligence. California: AAAI Press, 2008: 830-835.

［12］ Gabrilovich E, Markovitch S. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis［C］//Proc of The 20th International Joint Conference for Artificial Intelligence. California: AAAI Press, 2007:1606-1611.

［13］ Li Yanjun, Chung S M, Holt J D. Text Document Clustering Based on Frequent Word Meaning Sequences ［J］. Data ＆ Knowledge Engineering, 2008, 64(1): 381-404.

［14］ Shaban K. A Semantic Approach for Document Clustering ［J］. Journal of Software, 2009, 4(5): 391-404.

［15］ Gad W K, Kamel M S. New Semantic Similarity Based Model for Text Clustering Using Extended Gloss Overlaps［C］//Proc of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition. Berlin: Springer-Verlag, 2009: 663-677.

［16］ Liu Jianyi, Wang Jinghua, Wang Cong. Research on Text Network Representation ［C］//Proc of IEEE International Conference on Networking, Sensing and Control. Washington: IEEE Computer Society, 2008: 1217-1221.

［17］ 李益红, 卢朝阳, 李静, 等. 一种提取局部区域共同向量的瑕疵分类算法［J］. 西安电子科技大学学报, 2011, 38(5): 59-64.

Li Yihong, Lu Zhaoyang, Li Jing, et al. Algorithm for Extraction of the Local Region Common Vector for Defect Classification ［J］. Journal of Xidian University, 2011, 38(5): 59-64.

本刊中的类似文章

1．宁卓1;2;龚俭1;2.基于最大属性熵的GIDS报文分类算法
[J]. 西安电子科技大学学报, 2007,34(7): 201-204
2．王勇1;2;陶晓玲1.

分级结构的AdaBoost入侵检测方法研究

[J]. 西安电子科技大学学报, 2008,35(2): 345-350

3．暂时无作者信息.RBF网络在通信信号自动识别中的应用[J]. 西安电子科技大学学报, 1996,23(1):