现代图书情报技术 » 2013, Vol. 29 » Issue (9) :30-34

词语位置加权TextRank的关键词抽取研究

夏天[1,2]

1. 中国人民大学数据工程与知识工程教育部重点实验室 北京 100872;
2. 中国人民大学信息资源管理学院 北京100872

Xia Tian[1,2]

Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education, Renmin University of China, Beijing 100872, China) (School of Information Resource Management, Renmin University of China, Beijing 100872, China

摘要 把关键词抽取问题看作是构成文档词语的重要性排序问题,基于TextRank基本思想,构建候选关键词图,引入覆盖影响力、位置影响力和频度影响力用于计算词语之间的影响力概率转移矩阵,通过迭代法实现候选关键词分值计算,并挑选前N个作为关键词抽取结果。实验结果表明,对词语位置加权的TextRank方法优于传统的TextRank方法和基于LDA主题模型的关键词抽取方法。

关键词: 关键词抽取 词排序 TextRank 图模型 LDA

Abstract: The keyword extraction problem is taken as a word importance ranking problem. In this paper,candidate keyword graph is constructed based on TextRank, and the influences of word coverage, location and frequency are used to calculate the probability transition matrix, then, the word score is calculated by iterative method, and the top N candidate keywords are picked as the final results. Experimental results show that the proposed word position weighted TextRank method is better than the traditional TextRank method and LDA topic model method.

Keywords: Keyword extraction, Word rank, TextRank, Graph model, LDA

收稿日期: 2013-07-01;

基金资助:本文系国家社会科学基金项目“Web2.0环境下的网络舆情采集与分析”(项目编号:09CTQ027)和国家社会科学基金重大项目“云计算环境下的信息资源集成与服务研究”(项目编号:12&ZD220)的研究成果之一。

通讯作者 夏天 Email: xiat@ruc.edu.cn

引用本文:

夏天 .词语位置加权TextRank的关键词抽取研究[J] 现代图书情报技术, 2013,V29(9): 30-34

Xia Tian .Study on Keyword Extraction Using Word Position Weighted TextRank[J] , 2013,V29(9): 30-34

链接本文:

http://www.infotech.ac.cn/CN/ 或 http://www.infotech.ac.cn/CN/Y2013/V29/I9/30

[1] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[C]. In: *Proceedings of Empirical Methods in Natural Language Processing*, Barcelona, Spain. 2004:404-411.

[2] Frank E, Paynter G W, Witten I H, et al. Domain-Specific Keyphrase Extraction[C]. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden. 1999: 668-673.

[3] Turney P D. Learning Algorithms for Keyphrase Extraction[J]. *Information Retrieval*, 2000, 2(4):303-336.

[4] Pasquier C. Task 5: Single Document Keyphrase Extraction Using Sentence Clustering and Latent Dirichlet Allocation[C]. In: *Proceedings the 5th International Workshop on Semantic Evaluation.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2010: 154-157

[5] 石晶,李万龙. 基于LDA模型的主题词抽取方法[J]. 计算机工程, 2010, 36(19):81-83.(Shi Jing, Li Wanlong. Topic Words Extraction Method Based o LDA Model[J]. *Computer Engineering*, 2010, 36(19): 81-83.)

[6] 刘俊,邹东升,邢欣来,等. 基于主题特征的关键词抽取[J]. 计算机应用研究, 2012, 29(11): 4224-4227. (Liu Jun, Zou Dongsheng, Xing Xinlai, et al. Keyphrase Extraction Based on Topic Feature [J]. *Application Research of Computers*, 2012, 29(11): 4224-4227.)

[7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. *Journal of Machine Learning Research*, 2003, 3: 993-1022.

[8] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web [R]. Stanford Digital Library Technologies Project,1998.

[9] Rajaraman A, Ullman J D. Mining of Massive Datasets[M]. Cambridge University Press, 2012: 171-173.

[10] 夏天. 中心网页中主题网页链接的自动抽取[J]. 山东大学学报:理学版, 2012, 47(5): 25-31. (Xia Tian. Automatic Extracting Topic Page Links from H Page[J]. *Journal of Shandong University: Natural Science*, 2012, 47(5): 25-31.)

[11] 夏天. 基于扩展标记树的网页正文抽取[J]. 广西师范大学学报:自然科学版, 2011, 29(1): 133-137. (Xia Tian. Content Extraction of Web Page Basec Extended Label Tree[J]. *Journal of Guangxi Normal University: Natural Science Edition*, 2011, 29(1): 133-137.)

[1] 胡勇军, 江嘉欣, 常会友.基于LDA高频词扩展的中文短文本分类[J]. 现代图书情报技术, 2013,(6): 42-48

[2] 王嘉琦, 徐朝军, 李艺.基于LDA模型的社交网站自动量化评价研究[J]. 现代图书情报技术, 2013,29(3): 58-64

[3] 叶春蕾, 冷伏海.基于词汇链的路线图关键词抽取方法研究[J]. 现代图书情报技术, 2013,29(1): 50-56

[4] 范云满, 马建霞.利用LDA的领域新兴主题探测技术综述[J]. 现代图书情报技术, 2012,(12): 58-65

[5] 单斌, 李芳.基于种子文档LDA话题的演化研究[J]. 现代图书情报技术, 2011,27(7/8): 104-109

[6] 李欣, 刘丹.基于LDAP实现多认证源的统一身份认证实践——以华东师范大学图书馆为例[J]. 现代图书情报技术, 2011,27(4): 89-93

[7] 王昊, 邓三鸿, 苏新宁.基于字序列标注的中文关键词抽取研究[J]. 现代图书情报技术, 2011,27(12): 39-45

[8] 殷蜀梅,张智雄,吴振新.一种从医学文本中实现自动关键词抽取和筛选的技术方法*[J]. 现代图书情报技术, 2008,24(8): 31-36