现代图书情报技术 » 2014, Vol. 30 » Issue (1) :51-55

科技文献术语的自动抽取技术研究与分析

曾文, 徐硕, 张运良, 翟娟华

中国科学技术信息研究所 北京 100038

Zeng Wen, Xu Shuo, Zhang Yunliang, Zhai Juanhua

Institute of Scientific & Technical Information of China, Beijing 100038, China

- 摘要
- 参考文献
- 相关文章

Download: PDF (454KB) HTML (1KB) Export: BibTeX or EndNote (RIS) Supporting Info

摘要 【目的】为提高科技文献信息的组织和检索效率，从解决科技文献术语抽取这一基础研究问题入手，提出一种基于科技文献术语特点和统计计算相结合的科技文献术语自动抽取方法。【方法】核心技术是结合科技文献术语的语言特点，以及术语在文献中的词语组合强度和出现位置等统计计算信息，构建科技文献术语自动抽取算法。【结果】实验测试结果表明，获取的科技文献术语词语的平均准确率可以达到51.2%。【局限】在统计计算算法和数据处理方面，还需进一步改进算法和提高数据质量。【结论】提出的基于科技文献术语特点和统计计算相结合的科技文献术语自动抽取方法是有效的。

关键词： 科技术语 术语特点 统计计算 自动抽取

Abstract： [Objective] In order to improve the efficiency of science and technology literature information organization and retrieval,extraction of science and technology terms is the basic research problem. [Methods] The paper proposes an automatic extraction method based on science and technology terms characteristics and statistical computing. The method fully combines language characteristics and statistical information of terms such as the combination strength between words and the position that appeared in the literature to realize automatic extraction algorithm. [Results] Experimental results show that the average accuracy of scientific terms extraction can reach 51.2%. [Limitations] Statistical computing algorithm and data processing still need further improve for the algorithm and the quality of data. [Conclusions] The proposed method is effective.

Keywords： Technical term, Term characteristic, Statistical calculation, Automatic extraction

收稿日期: 2013-08-15;

基金资助:

本文系“十二五”国家科技支撑计划课题“基于多源信息的电动汽车数据挖掘关键技术研究”（项目编号：2013BAG06B01）和国家自然科学基金项目“支持面向特定情报分析应用的知识组织系统快速构建关键问题研究”（项目编号：71203208）的研究成果之一。

通讯作者 曾文 Email: zengw@istic.ac.cn

作者贡献: 曾文：提出研究思路，负责设计研究方案和进行实验分析，撰写论文； 徐硕：数据采集； 张运良：数据清洗；翟娟华：论文修订。

引用本文:

曾文, 徐硕, 张运良等 .科技文献术语的自动抽取技术研究与分析[J] 现代图书情报技术, 2014,V30(1): 51-55

Zeng Wen, Xu Shuo, Zhang Yunliang etc .The Research and Analysis on Automatic Extraction of Science and Technology Literature Terms[J] 2014,V30(1): 51-55

链接本文:

http://www.infotech.ac.cn/CN/ 或 http://www.infotech.ac.cn/CN/Y2014/V30/I1/51

[1] Frantzi K T,Ananiadou S,Mima H. Automatic Recognition of Multi-word Terms：The C-value/NC-value Method [J]. International Journal on Digital Libraries,2000,3（2）：115-130.

[2] 常鹏,马辉.高效的短文本主题词抽取方法[J]. 计算机工程与应用,2011,47（20）：126-128,154.（Chang Peng,Ma Hui. Efficient Short Texts Keyword Extraction Method Analysis[J]. Computer Engineering and Applications,2011,47（20）：126-128,154.）

[3] 李鹏,王斌,石志伟,等. Tag-TextRank：一种基于Tag的网页关键词抽取方法[J]. 计算机研究与发展,2012,49（11）：2344-2351.（Li Peng,Wang Bin,Shi

Service
把本文推荐给朋友
加入我的书架
加入引用管理器
Email Alert
RSS
作者相关文章
曾文
徐硕
张运良
翟娟华

Zhiwei,et al. Tag-TextRank：A Webpage Keyword Extraction Method Based on Tags[J]. Journal of Computer Research and Development,2012,49（11）：2344-2351.）

[4] 陈文亮,朱靖波,姚天顺,等. 基于Bootstrapping的领域词汇自动获取[C]. 见：全国第7届计算语言学联合学术会议论文集.2003：67-72.（Chen Wenliang,Zhu Jingbo,Yao Tianshun,et al. Automatic Learning Field Words by Bootstrapping[C]. In：Proceedings of the 7th Computational Linguistics in China. 2003：67-72.）

[5] 王裴岩,张桂平,蔡东风,等. 一种用于专利主题词抽取的模板自动生成方法[J]. 沈阳航空工业学院学报,2010,27（3）：46-49.（Wang Peiyan ,Zhang Guiping,Cai Dongfeng,et al. An Automation Pattern Generation Method for Patent Topic Keyword Extraction[J]. Journal of Shenyang Institute of Aeronautical Engineering,2010,27（3）：46-49.）

[6] 邢红兵. 信息领域汉语术语的特征及其在语料中的分布规律[J]. 术语标准化与信息技术,2000（3）：17-21.（Xing Hongbing. Structural Features and Distributions of Chinese- English Terms in the Corpus from Information Field[J]. Terminology Standardization and Information Technology,2（3）：17-21.）

[7] 张榕. 术语定义抽取、聚类与术语识别研究[D]. 北京：北京语言大学,2006.（Zhang Rong. The Term Definition Extraction, Clustering and Terminology Recognition Research [D]. Beijing：Beijing Language and Culture University,2006.

[1] 张秀秀,马建霞.PDF科技论文语义元数据的自动抽取研究*[J]. 现代图书情报技术, 2009,3(2): 102-106

[2] 曾苏,马建霞,张秀秀.元数据自动抽取研究新进展*[J]. 现代图书情报技术, 2008,24(4): 7-11

[3] 何琳.领域本体的关系抽取研究*[J]. 现代图书情报技术, 2008,24(4): 35-38

[4] 谈春梅,颜世伟,刘子牧.网络专题知识组织知识元自动抽取系统的设计与实现*[J]. 现代图书情报技术, 2008,24(3): 62-67

[5] 王璐,朱东华,仁智军 .科技术语属性抽取方法研究*[J]. 现代图书情报技术, 2007,2(5): 69-72

[6] 王永成.自动编制文献摘要及知识的自动提取[J]. 现代图书情报技术, 1993,9(3): 13-13