基于自主学习规则的中文物种描述文本的语义标注研究

段宇锋[1], 黑珍珍[1], 鞠菲[1], 崔红[2]

1. 华东师范大学商学院 上海 200241;
2. 美国亚利桑那大学图书馆学与信息资源学院 图森 85719

Duan Yufeng[1], Hei Zhenzhen[1], Ju Fei[1], Cui Hong[2]

1. Business School, East China Normal University, Shanghai 200241, China;
2. School of Information Resource & Library Science, University of Arizona, Tucson 85719, USA

摘要 从《中国植物志》中随机采集1 000个文档作为数据集,采用自主学习规则与先导词相结合的算法实现中文物种描述文本的语义标注。实验数据表明,本研究设计的基于规则的算法整体标注效率(F值)达到0.930,大部分元素的F值在0.724-0.964之间,该算法优于朴素贝叶斯分类算法。同时证明,先导词对优化算法具有积极意义。

关键词: 规则 先导词 物种描述文本 语义标注

Abstract: This paper uses the algorithm of auto-learning rules combining with leading words to implement the semantic markup of species description text in Chinese with the data set of 1 000 documents collected from Flora of China randomly. Experimental results indicate that the whole markup efficiency (the values of F) of rule-based algorithm, which is designed by the study, generally reaches 0.930, and most elements are in the range of 0.724-0.964. Therefore, this algorithm is better than Naive Bayesian categorization algorithm, and it is also proved that leading words are positive for optimizing the algorithm.

Keywords: Rules, Leading words, Species description text, Semantic markup

收稿日期: 2012-03-26;

基金资助:

本文系教育部人文社会科学青年项目“基于深度语义标注的网络中文学术信息抽取研究”(项目编号:10YJC870004)的研究成果之一。

引用本文:

段宇锋, 黑珍珍, 鞠菲等 .基于自主学习规则的中文物种描述文本的语义标注研究[J] 现代图书情报技术, 2012,V28(5): 41-47

Duan Yufeng, Hei Zhenzhen, Ju Fei etc .Study on Semantic Markup of Species Description Text in Chinese Based on Auto-learning Rules[J] , 2012,V28(5): 41-47

链接本文:

http://www.infotech.ac.cn/CN/ 或 http://www.infotech.ac.cn/CN/Y2012/V28/I5/41

[1] Taylor A. Extracting Knowledge from Biological Descriptions[C]. In: *Proceedings of the 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases*. 1995:114-119.

[2] Vanel J M. Worldwide Botanical Knowledge Base[EB/OL]. [2011-10-11]. http://wwbota.free.fr/.

[3] Wood M M, Lydon S J, Tablan V, et al. Using Parallel Texts to Improve Recall in IE[C]. In: *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP)*.Amsterdam: John Benjamins, 2004:70-77.

[4] 罗贝,吴洁,曹存根,等. 从文本中获取植物知识方法的研究[J]. 计算机科学 ,2005,32(10):6-13.(Luo Bei, Wu Jie, Cao Cungen,et al. Botanical Knowledge Acquisition from Text[J]. *Computer Science*, 2005,32(10):6-13.)

[5] 沙丽华. 面向领域文档的语义标注方法研究[D]. 长春:吉林大学,2009.(Sha Lihua. Research on Semantic Annotation for Domain Documents[D]. Changchun: Jilin University,2009.)

[6] 石静. 基于本体的植物信息抽取与分析研究[D]. 西安:西北农林科技大学,2010. (Shi Jing. Information Extraction and Analysis Based on Plant Ontology [D]. Xi'an: Northwest Agriculture and Foresty University, 2010.)

[7] Sautter G, Bohm K, Agosti D. A Combining Approach to Find all Taxon Names[J]. *Biodiversity Informatics*,2006(3):46-58.

[8] Tang X Y, Heidorn P B. Using Automatically Extracted Information in Species Page Retrieval[EB/OL]. [2011-08-10]. http://www.tdwg.org/proceedings/article/view/195/.

[9] Soderland S. Learning Information Extraction Rules for Semi-Structured and Free Text[J]. *Machine Learning*, 1999, 34 (1-3): 233-272.

[10] 郑家恒,菅小艳. 农作物信息抽取系统的设计与实现[J]. 计算机工程 ,2006,32(7):197-198,220.(Zheng Jiaheng, Jian Xiaoyan. Design and Realization of the System of Farm Crop Information Extraction[J]. *Computer Engineering*, 2006, 32(7):197-198,220.)

[11] Cui H, Heidorn P B. The Reusability of Induced Knowledge for Automatic Semantic Markup of Taxonomic Descriptions[J]. *Journal of the American Society for Information Science and Technology*, 2007, 58(1): 133-149.

[12] Cui H, Boufford D, Selden P. Semantic Annotation of Biosystematics Literature Without Training Examples[J]. *Journal of the American Society of Information Science and Technology*, 2010, 61 (3): 522-542.

[13] Cui H. The XML Schema for MARTT[EB/OL]. [2012-08-08]. http://publish.uwo.ca/~hcui7/research/xmlschema.xsd.

[14] 中国植物志编辑委员会. 中国植物志[M]. 北京:科学出版社,1959.(Flora of China Editorial Committee. Flora of China [M]. Beijing: Science Press, 1959.)

[1] 刘萍, 胡月红.基于FCA和关联规则的情报学本体构建[J]. 现代图书情报技术, 2012,28(2): 34-40

[2] 黄名选, 余如.基于负关联规则与频繁项集挖掘的信息检索系统[J]. 现代图书情报技术, 2011,27(7/8): 91-96

[3] 胡元蛟, 王昊.面向CSSCI的学者知识地图构建与分析[J]. 现代图书情报技术, 2011,27(3): 38-44

[4] 丁晟春, 江超男.基于SWRL规则推理的隐含关系挖掘[J]. 现代图书情报技术, 2011,27(3): 68-72

[5] 路永和, 曹利朝.基于关联规则综合评价的图书推荐模型[J]. 现代图书情报技术, 2011,27(2): 81-86

[6] 梁文超, 徐朝军, 沈书生.模糊规则算法在教育信息分类中的应用[J]. 现代图书情报技术, 2011,27(1): 94-98

[7] 牟冬梅, 范轶, 王丽伟.数字资源语义互联研究(III) ——语义标注子系统的设计与实现[J]. 现代图书情报技术, 2010,26(9): 13-17

[8] 陈瑗瑛 秦宗蓉.基于FP-tree的中小馆书目数据库主题词数据挖掘*[J]. 现代图书情报技术, 2010,26(7/8): 114-119

[9] 李楠 郑荣廷 吉久明 滕青青.基于启发式规则的中文化学物质命名识别研究[J]. 现代图书情报技术, 2010,26(5): 13-17

[10] 滕广青,毕强.基于概念格的数字图书馆用户用法细分* ——数字图书馆用户使用方法的关联规则挖掘[J]. 现代图书情报技术, 2010,26(3): 8-12

[11] 王昊, 苏新宁.基于CSSCI本体的学科关联分析[J]. 现代图书情报技术, 2010,26(10): 10-16

[12] 滕广青, 毕强.概念格构建工具ConExp与Lattice Miner的比较研究[J]. 现代图书情报技术, 2010,26(10): 17-22

[13] 夏彦, 何琳, 潘运来, 欧阳辰晨.基于规则与统计相结合的互联网突发事件识别研究[J]. 现代图书情报技术, 2010,26(10): 65-69

[14] 陈欣,李晓菲.基于领域本体的专业文献信息检索研究[J]. 现代图书情报技术, 2009,25(7-8): 59-64

[15] 施聪莺,徐朝军,杨晓江.基于规则和Rocchio分类器的学前综合教育资源分类*[J]. 现代图书情报技术, 2009,25(7-8): 75-79