现代图书情报技术 » 2014, Vol. 30 » Issue (1) :43-50

领域本体术语抽取研究

汤青[1], 吕学强[1,2], 李卓[1], 施水才[1,2]

1 北京信息科技大学网络文化与数字传播北京市重点实验室 北京 100101;
2 北京拓尔思信息技术股份有限公司 北京 100101

Tang Qing[1], Lv Xueqiang[1,2], Li Zhuo[1], Shi Shuicai[1,2]

1 Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China;
2 Beijing TRS Information Technology Co.Ltd., Beijing 100101, China

- 摘要
- 参考文献
- 相关文章

Download: PDF (608KB) HTML (1KB) Export: BibTeX or EndNote (RIS) Supporting Info

摘要 【目的】尽可能多地抽取多字词本体术语，以保证本体构建的质量。【方法】提出基于部件扩展的本体术语抽取方法。利用部件的领域聚合性和词性特征，采用领域词频比较的方法抽取部件；考虑术语长度、术语词性构成以及术语内部结合度等因素，设计合理的扩展规则对部件扩展以形成候选术语；利用上下文关联信息、语境信息从候选术语集中筛选出本体术语。【结果】利用该方法在IT领域实验数据集上进行测试，实验结果准确率为83.5%，召回率为87%，准确率相比Baseline方法要高出2.5个百分点。【局限】部件抽取方法需要借助于平衡语料库，部件的质量直接影响术语抽取效果。【结论】实验结果表明该方法是有效的，对本体学习、本体构建具有积极意义。

关键词： 本体术语 术语抽取 术语部件 部件扩展

Abstract： [Objective] Ontology terms are extracted as more as possible for the quality of Ontology construction. [Methods] This paper proposes an Ontology term extraction method based on term component extension. It uses the polymerization characteristics and POS features of the terms,extracts term components by word frequency comparison approach. Considering the factors of term length,term POS and term internal associative strength of character strings,reasonable extended rules are designed for components extension to get the candidate terms. Then,Ontology terms are filtered from candidate terms by using the relational information and the contextual information. [Results] Experimental result shows that accuracy rate is 83.5%,the recall rate is 87%,the accuracy rate is 2.5 percentages over the baseline. [Limitations] It needs a balanced corpus to extract term component,and term extracting effect is effected by the quality of the term. [Conclusions] The method is effective and has a positive significance for Ontology learning and Ontology construction etc.

Keywords： Ontology term, Term extraction, Term component, Component extension

收稿日期: 2013-09-27;

基金资助:

本文系国家自然科学基金项目“基于本体的专利自动标引研究”（项目编号：61271304）和北京市教委科技发展计划重点项目暨北京市自然科学基金B类重点项目“面向领域的互联网多模态信息精准搜索方法研究”（项目编号：KZ201311232037）的研究成果之一。

通讯作者 汤青 Email: tangqing20062008@126.com

作者贡献: 汤青：提出研究思路，设计研究方案和完成实验，论文的起草、撰写； 吕学强，李卓：负责设计论文框架和论文的修改；施水才：提出研究课题，负责论文的修订工作。

引用本文:

汤青, 吕学强, 李卓等 .领域本体术语抽取研究[J] 现代图书情报技术, 2014,V30(1): 43-50

Tang Qing, Lv Xueqiang, Li Zhuo etc .Research on Domain Ontology Term Extraction[J] , 2014,V30(1): 43-50

链接本文:

http://www.infotech.ac.cn/CN/ 或 http://www.infotech.ac.cn/CN/Y2014/V30/I1/43

[1] Gruber T R. A Translation Approach to Portable Ontology Specifications [J]. Knowledge Acquisition,1993,5（2）：199-220.

[2] Chambers N,Jurafsky D. Template-based Information Extraction without the Templates [C]. In：Proceedings of the 49th Annual Meeting the Association for Computational Linguistics（HLT’11）. Stroudsburg：Association for Comp- utational Linguistics,2011：976-9

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
汤青
吕学强
李卓
施水才

[3] 韦小丽,孙涌,张书奎,等.基于最大熵模型的本体概念获取方法研究 [J]. 计算机工程,2009,35（24）：114-116.（Wei Xiaoli,Sun Yong,Zhang Shukui,et a Ontological Concept Extraction Method Based on Maximum Entropy Model [J]. Computer Engineering,2009,35（24）：114-116.）

[4] 游宏梁,张巍,沈钧毅,等.一种基于加权投票的术语自动识别方法[J]. 中文信息学报,2011,25（3）：9-16.（You Hongliang,Zhang Wei,Shen Junyi,et al. Weighted Voting Based Automatic Term Recognition Method[J]. Journal of Chinese Information Processing,2011,25（3）：9-16.）

[5] Yang Y,Lu Q,Zhao T.A Delimiter-based General Approach for Chinese Term Extraction [J]. Journal of the American Society for Informatior Science and Technology,2010,61（1）：111-125.

[6] 章成志.基于多层术语度的一体化术语抽取研究[J]. 情报学报,2011,30（3）：275-285.（Zhang Chengzhi.Using Integration Strategy and Multi-level Termhood to Extract Terminology [J]. Journal of the China Society for Scientific and Technical Information,2011,30（3）：275-285.）

[7] Lee C,Huang C,Tang K,et al. Iterative Machine-Learning Chinese Term Extraction [C]. In：Proceedings of the 14th International Conferer on Asia-Pacific Digital Libraries. 2012：309-312.

[8] 王卫民,贺冬春,符建辉.基于种子扩充的专业术语识别方法研究[J]. 计算机应用研究,2012,29（11）：4105-4107.（Wang Weimin,He Dongchun,Fu Jianh Research of Professional Term Identification Method Based on Seed Expansion[J]. Application Research of Computers,2012,29（11）：410 4107.）

[9] 吴云芳,穗志方,邱利坤,等.信息科学与技术领域术语部件描述[J]. 语言文字应用,2003（4）：34-39.（Wu Yunfang,Sui Zhifang,Qiu Likun,et al. The Approaches and Strategies to Describe the Term Component in Information Science and Technology [J]. Applied Linguistics,2003（4）：3 39.）

[10] 冯志伟. 术语形成的经济律——FEL公式[J]. 中国科技术语,2010,12（2）：9-15.（Feng Zhiwei.Economic Law of Term Formation——FEL Formula [J]. China Terminology,2010,12（2）：9-15.）

[11] 李萍,黄崇岭. IT领域的专业术语构词特点及功能意义[J].桂林电子工业学院学报,2004,24（2）：48-51.（Li Ping,Huang Chongling. The Morphological Formation and Functional Significance of Technical Term in IT Field [J]. Journal of Guilin University of Electronic Technology,2004,24（2） 48-51.）

[12] 陈士超,郁滨. 面向术语抽取的双阈值互信息过滤方法[J].计算机应用,2011,31（4）：1070-1073.（Chen Shichao,Yu Bin. Method of Mutual Informati Filtration with Dual-threshold for Term Extraction[J]. Journal of Computer Applications,2011,31（4）：1070-1073.）

[13] Page L,Brin S,Motwani R,et al. The PageRank Citation Ranking：Bringing Order to the Web[R]. Stanford InfoLab,1999.

[14] Resnik P. Using Information Content to Evaluate Semantic Similarity [C]. In：Proceedings of the 14th International Joint Conference on Artificial Intelligence（IJCAI’95）. San Francisco：Morgan Kaufmann Publishers Inc.,1995：448- 453.

[15] Tan P,Steinbach M,Kumar V. Introduction to Data Mining [M]. Addison-Wesley,2005.

[16] 何琳.基于多策略的领域本体术语抽取研究[J]. 情报学报,2012,31（8）：798-804.（He Lin. Domain Ontology Terminology Extraction Based on Integrated Strategy Method [J]. Journal of the China Society for Scientific and Technical Information,2012,31（8）：798-804.）

[1] 熊李艳, 谭龙, 钟茂生.基于有效词频的改进C-value自动术语抽取方法[J]. 现代图书情报技术, 2013,29(9): 54-59

[2] 化柏林.针对中文学术文献的情报方法术语抽取[J]. 现代图书情报技术, 2013,(6): 68-75

[3] 胡阿沛, 张静, 刘俊丽.基于改进C-value方法的中文术语抽取[J]. 现代图书情报技术, 2013,29(2): 24-29

[4] 李振清, 刘建毅, 王枞, 吴旭.同行评议专家遴选系统研究与实现[J]. 现代图书情报技术, 2012,28(5): 81-86

[5] 康小丽, 章成志.用于双语术语抽取的专业领域中英文可比语料库构建[J]. 现代图书情报技术, 2012,28(2): 28-33