基于FCA和关联规则的情报学本体构建

刘萍, 胡月红

武汉大学信息资源研究中心 武汉 430072

Liu Ping, Hu Yuehong

Center for Studies of Information Resources, Wuhan University, Wuhan 430072, China

- 摘要
- 参考文献
- 相关文章

Download: PDF (840KB) HTML (1KB) Export: BibTeX or EndNote (RIS) Supporting Info

摘要 提出一种新的领域本体学习方法,结合形式概念分析(FCA)与关联规则挖掘从非结构化文本中获取情报学本体。该方法从文本集中通过种子-扩展机制的方法获取领域核心概念,构建文档概念格(文档×关键词矩阵),在此基础上通过形式概念分析方法来识别概念之间的等级关系,通过关联规则挖掘概念间的相关关系。最后,采用基于“黄金标准”的方法对本体学习的结果进行评价,结果表明:通过这种方法构建的本体可以达到较高的领域知识覆盖率,而且能够识别概念之间部分隐含的关系,从而验证该方法在领域本体的构建中实用且有效。

关键词: 本体构建 情报学 FCA 关联规则

Abstract: This paper presents a new approach to Ontology learning in the domain of information science. A combination of Formal Concept Analysis (FCA) and association rules is used to facilitate Ontology construction from unstructured text. This approach acquires key concepts from documents by using a seeding and expansion mechanism; formulates (key concept by document) context for concept lattice construction, and bootstraps the learning of domain-specific concept hierarchies using FCA; extracts the relationships between the concepts via association rules. To evaluate the quality of the learned Ontology, a comparison with “Golden Standard” is undertaken, and the evaluation results illustrate that it can reach high domain coverage and identify some implicit relations between concepts. It is concluded that the proposed method is practical and useful to support the process of building domain Ontology.

Keywords: Ontology development, Information science, FCA, Association rule

收稿日期: 2011-09-08;

基金资助:

本文系教育部人文社会科学研究青年基金项目“高校专家知识地图构建研究”(项目编号:10YJC870022)的研究成果之一。

引用本文:

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,V28(2): 34-40

Liu Ping, Hu Yuehong .Development of Domain Ontology in Information Science Based on FCA and Association Rules[J] , 2012,V28(2): 34-40

链接本文:

http://www.infotech.ac.cn/CN/ 或 http://www.infotech.ac.cn/CN/Y2012/V28/I2/34

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
- 刘萍
- 胡月红

[1] Braschler M, Schäuble P. Multilingual Information Retrieval Based on Document Alignment Techniques [C]. In: *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*. London, UK:Springer-Verlag,1998:183-197.

[2] Gu T. Using Formal Concept Analysis for Ontology Structuring and Building[C]. In: *Proceedings of the International Standard Industrial Classification(ISIC)*. Nanyang Technological University, 2003.

[3] 孙茂松.汉语自动分词研究中的若干理论问题[J]. 语言文字应用 ,2005 (4):40-46. (Sun Maosong.Several Theoretical Problems in Automatic Chinese Word Segmentation Research[J]. *Application of Language*,2005(4):40-46.)

[4] 张德鑫. “水至清则无鱼”——我的新生词语规范观[J]. 北京大学学报:哲社版 ,2000,37(5):106-119.(Zhang Dexin. My Point of View on the Standard of Newborn Words[J]. *Journal of Peking University:Philosophy and Social Sciences*,2000,37(5):106-119.)

[5] 黄昌宁,赵海.中文分词十年回顾[J]. 中文信息学报 ,2007,21(3):8-19.(Huang Changning, Zhao Hai. Chinese Word Segmentation: A Decade Review

[J]. *Journal of Chinese Information Processing*,2007,21(3):8-19.)

[6] Haav H M. A Semi-automatic Method to Ontology Design by Using FCA[C]. In: *Proceedings of the 2nd International CLA Workshop, Concept Lattices and Their Applications*. Technical University of Ostrava, 2004: 13-25.

[7] Leturia I, Vicente I S,Saralegi X. Search Engine Based Approaches for Collecting Domain-Specific Basque-English Comparable Corpora from the Internet [C].In: *Proceedings of the 5th Web as Corpus Workshop (WAC5)*. San Sebastian: Elhuyar Fundazioa, 2009:53-61.

[8] 张海军,史树敏,朱朝勇,等.中文新词识别技术综述[J]. 计算机科学 ,2010,37(3): 6-12.(Zhang Haijun, Shi Shumin, Zhu Chaoyong,et al. Survey of Chinese New Words Identification[J]. *Computer Science*,2010,37(3):6-12.)

[9] 郑家恒,李文花.基于构词法的网络新词自动识别初探[J]. 山西大学学报:自然科学版 ,2002,25(2):115-119.(Zhen Jiaheng, Li Wenhua. Study on Automatic Identification for Internet New Words According to Word-Building Rule[J]. *Journal of Shanxi University:Natural Science Edition*,2002,25 (2):115-119.)

[10] Chen K J, Bai M H. Unknown Word Detection for Chinese by a Corpus-based Learning Method[J]. *International Journal of Computational Linguistics and Chinese Language Processing*,1998, 3(1):27-44.

[11] 吴涛,张毛迪,陈传波.一种改进的统计与后串最大匹配的中文分词算法研究[J]. 计算机工程与科学 ,2008,30(8):79-82.(Wu Tao, Zhang Maodi, Chen Chuanbo. Research of Chinese Word Segmentation Algorithms Based on Statistics and Reverse Maximum Match[J]. *Computer Engineering & Science*,2008,30(8):79-82.)

[12] Nie J Y, Hannah M L, Jin W. Unknown Word Detection and Segmentation of Chinese Using Statistical and Heuristic Knowledge [J].*Communications of COLIPS*,1995,5(1):47-57.

[13] Obitko M, Sná el V, Smid J. Ontology Design with Formal Concept Analysis [C]. In: *Proceedings of the CLA 2004 International Workshop on Concept Lattices and Their Applications*. Technical University of Ostrava, 2004: 111-119.

[14] Pirkola A,Leppanen E,Jarvelin K.The RATF Formula (Kwok's Formula): Exploiting Average Term Frequency in Cross-Language Retrieval[J/OL]. *Information Research*,2002,7(2). [2010-01-05].http://InformationR.net/ir/7-2/infres72.html.

[15] Han J, Kamber M. Data Mining: Concepts and Techniques [R/OL]. [2011-01-23]. http://134.208.3.165/course/2006/Fall/Data_mining/06.pdf.

[16] Yang M H, Ahuja N. A Geometric Approach to Train Support Vector Machines[C]. In:*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, SC, USA.2000:430-437.

[17] 秦浩伟,步丰林.一个中文新词识别特征的研究[J]. 计算机工程 ,2004,30(S1):369-370.(Qin Haowei, Bu Fenglin. Research on a Feature of Chinese New Word Identification[J].*Computer Engineering*,2004,30(S1):369-379.)

[18] 李钝,曹元大,万月亮.Internet中的新词识别[J]. 北京邮电大学学报 ,2008,31(1):26-29.(Li Dun, Cao Yuanda, Wan Yueliang. Internet-Oriented New Words Identification[J]. *Journal of Beijing University of Posts and Telecommunications*,2008,31(1):26-29.)

[19] 韩艳,林煜熙,姚建民. 基于统计信息的未登录词的扩展识别方法[J]. 中文信息学报 ,2009,23(3): 24-30.(Han Yan, Lin Yixi, Yao Jianmin. Study on Chinese OOV Identification Based on Extension[J].*Journal of Chinese Information Processing*,2009,23(3):24-30.)

[20] 丁建立,慈祥,黄剑雄.一种基于免疫遗传算法的网络新词识别方法[J]. 计算机科学 ,2011,38(1): 240-245.(Ding Jianli, Ci Xiang, Huang Jianxiong. Approach of Internet New Word Identification Based on Innmune Genetic Algorithm[J].*Computer Science*, 2011,38(1):240-245.)

[21] Maedche A, Staab S. Discovering Conceptual Relations from Text[C]. In: *Proceedings of the European Conference on Artificial Intelligence (ECAI)*. Amsterdam:IOS Press, 2000: 321-325.

[22] 崔世起.中文新词检测与分析[D].北京:中国科学院研究生院,2006.(Cui Shiqi. Research on Chinese New Word Identification and Analysis [D]. Beijing: Graduate University of Chinese Academy of Sciences,2006.)

[23] Maedche A, Staab S. Ontology Learning for the Semantic Web[C]. In: *Proceedings of the IEEE Intelligent Systems*. 2001: 72-79.

[24] 韩客松,王永成,陈桂林.汉语语言的无词典分词模型系统[J]. 计算机应用研究 ,1999(10):8-9.(Han Kesong, Wang Yongcheng, Chen Guilin. Chinese Word Segmentation System Without Dictionary[J]. *Application Research of Computers*, 1999(10):8-9.)

[25] Keskustalo H, Hedlund T, Airio E. UTACLIR-General Query Translation Framework for Several Language Pairs[C]. In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York : Association for Computing Machinery, 2002:448.

[26] Lemur.The Lemur Toolkit for Language Modeling and Information Retrieval [EB/OL]. (2009-12-21).[2010-01-05].http://www.lemurproject.org/.

[27] 魏莎莎.一种中文未登录词识别及词典设计新方法[D].重庆:西南大学,2011.(Wei Shasha. A New Method of Chinese Out-of-Vocabulary Identification and Dictionary Design[D]. Chongqing: Southeast University,2011.)

[28] Noy N F, McGuinness D L. Ontology Development 101: A Guide to Creating Your First Ontology[R]. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001:4-23.

[29] Collier N, Kumano A, Hirakawa H. An Application of Local Relevance Feedback for Building Comparable Corpora from News Article Matching [J]. *Natl Inst Inform*, 2003(5):9-23.

[30] 贺敏.面向互联网的中文有意义串挖掘[D].北京:中国科学院研究生院,2007.(He Min. Internet-Oriented Chinese Meaningful Word Reorganization [D].Beijing: Graduate University of Chinese Academy of Sciences,2009.)

[31] Uschold M, Gruninger M. Ontologies: Principles, Methods and Applications[J]. *Knowledge Engneer Revision*,1996,11 (2):93-155.

[32] Rogati M, Yang Y M.CMU PRF Using a Comparable Corpus: CLEF Working Notes [C]. In: *Proceedings of Working Notes for the Cross-Language Evaluation Forum(CLEF 2001) Workshop*. Berlin:Springer-Verlag, 2001:81-86.

[33] Layiosa-Braithwaits S. Ensino Das Linguas Vivas no Superior em Portugal [M].Porto: Faculdade de Letras da Universidade do Porto,1999:307-317.

[34] 黄玉兰.有意义串挖掘及其应用[D].北京:中国科学院研究生院,2009.(Huang Yulan. Meaningful Word Reorganization and Application [D]. Beijing: Graduate University of Chinese Academy of Sciences,2009.)

[35] 贺敏,龚才春,张华平,等.一种基于大规模语料的新词识别方法[J]. 计算机工程与应用 ,2007,43(21):157-159.(He Min, Gong Caichun, Zhang Huaping,et al. Method of New Word of Identification Based on Lager-scale Corpus[J]. *Computer Engineering and Applications*,2007,43(21):157-159.)

[36] 张海军,史树敏,丁溪源,等.基于分词提取重复串的未登录词遗漏量化模型[J]. 中文信息学报 ,2011,25(2):122-128.(Zhang Haijun, Shi Shumin, Ding Xiyuan,et al. Quantitative Omission Model of Candidate Unknown Words for Chinese Word Segmentation Based Repeat Extraction[J].*Journal of Chinese Information Processing,*2011,25(2):122-128.)

[37] 何琳. 领域本体的半自动构建及检索研究[M]. 南京:东南大学出版社, 2009:99-100.(He Lin. Research on Semi-automatic Construction and Retrieval of Domain Ontology[M]. Nanjing: Southeast University Press, 2009:99-100.)

[38] Talvensaari T, Pirkola A, Jaervelin K, et al. Focused Web Crawling in the Acquisition of Comparable Corpora [J]. *Information Retrieval*,2008,11(5):427-445.

[39] Ji D H, Zhao S J, Xiao G Z. Chinese Document Re-ranking Based on Automatically Acquired Term Resource[J]. *Language Resource & Evaluation*, 2009, 43(4):385-406.

[40] 中国植物志[R/OL].[2011-09-12].http://frps.plantphoto.cn/dzb_list2.asp.(Flora of China [R/OL].[2011-09-12].http://frps.plantphoto.cn/dzb_list2.asp.)

[41] Yang Y M, Pederson J O. A Comparative Study on Feature Selection in Text Categorization[C]. In: *Proceedings of the 14th International Conference on Machine Learning.* Nashville:Morgan Kaufmann, 1997:412-420.

[42] Rogati M, Yang Y M.High-Performing Feature Selection for Text Classification[C]. In: *Proceedings of the 11th International Conference on Information and Knowledge Management*. New York:ACM,2002:659-661.

[43] Baroni M, Bernardini S. BootCaT: Bootstrapping Corpora and Terms from the Web[C]. In: *Proceedings of International Conference on Language Resources and Evaluation(LREC2004)*. Paris:European Language Resources Association, 2004:1313-1316.

[44] 陈小荷.自动分词中未登录词问题的一揽子解决方案[J]. 语言文字应用 ,1999(3):103-109.(Chen Xiaohe. A Package Scheme for Identifying Unlisted Words in Chinese Segmentation[J].*Applied Linguistics*,1999(3):103-109.)

[45] 都菁,熊海灵.基于论坛语料识别中文未登录词的方法[J]. 计算机工程与设计 ,2010,31(3):630-633.(Du Jing, Xiong Hailing. Algorithm to Recognize Unknown Chinese Words Based on BBS Corpus[J].*Computer Engineering and Design*,2010,31(3):630-633.)

[46] 吕美香,何琳, 李玥,等.基于N-gram文本表达的新闻领域关键词词典构建研究[J]. 情报科学 ,2010,28(4):571-574.(Lv Meixiang, He Lin, Li Yue,et al. Research on Construction of News Keyword Dictionary Based on N-Gram Text Representation[J]. *Intelligence Science*, 2010,28(4):571-574.)

[47] 王俊华. 基于文本的半监督领域本体构建[D]. 长春: 吉林大学, 2010:19-45.(Wang Junhua. Semi-supervised Domain Ontology Building Based on Text [D]. Changchun: Jilin University, 2010:19-45.)

[48] Overview on ConExp[EB/OL]. [2011-06-28]. http: //conexp.sourceforge.net /users.

[49] Rapp R.Identifying Word Translations in Non-parallel Texts[C].In: *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA:Association for Computational Linguistics,1995:320-322.

[50] IHMC CmapTools [EB/OL]. [2011-03-18]. http://cmap.ihmc.us/download/.

[51] 杜小勇,马文峰.领域本体评价研究[J]. 图书情报工作 ,2006,50(10):68-72.(Du Xiaoyong, Ma Wenfeng. An Evaluation Framework for Domain Ontology [J]. *Library and Information Service*, 2006,50(10):68-72.)

[52] 何琳.领域本体评价研究[J]. 图书馆杂志 ,2010,29(2):57-62.(He Lin. Research on Evaluation Mechanism of Domain Ontology[J]. *Library Journal*, 2010,29(2):57-62.)

[53] Bates M J. An Operational Definition of the Information Disciplines [EB/OL]. [2011-03-20].http://gseis.ucla.edu/faculty/bates/articles/pdf/Contribution512-1.pdf.

[54] Tanaka K, Iwasaki H.Extraction of Lexical Translations from Non-aligned Corpora[C]. In: *Proceedings of the 16th International Conference on Computational Linguistics*. Stroudsburg, PA:Association for Computational Linguistics,1996: 580-585.

[55] Sabou M, Wroe C, Goble C, et al. Learning Domain Ontologies for Web Service Descriptions: An Experiment in Bioinformatics[C]. In: *Proceedings of the 14th International World Wide Web Conference Committee (WWW2005)*. New York: ACM Press, 2005:190-198.

[1] 杨锐, 汤怡洁, 刘毅, 李崴.**Web环境中的本体构建系统研究综析**[J]. 现代图书情报技术, 2012,28(1): 13-18

[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] 滕广青, 毕强.**基于概念格的异构资源领域本体构建研究**[J]. 现代图书情报技术, 2011,27(5): 7-12

[4] 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

[5] 张云中.**一种基于FCA和Folksonomy的本体构建方法**[J]. 现代图书情报技术, 2011,27(12): 15-23

[6] 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

[7] 吴红 李玉平 胡泽文.**基于领域本体的专利信息检索系统研究与实现**[J]. 现代图书情报技术, 2010,26(6): 71-77
[8] 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
[9] 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
[10] 葛登科,王亚民.**基于GIS的空间关联规则挖掘方法研究**[J]. 现代图书情报技术, 2009,25(7-8): 97-101
[11] 陈亦佳,赵星.**基于期刊引文网络视角研究国际图书馆学情报学知识交流**[J]. 现代图书情报技术, 2009,25(6): 55-60
[12] 陈祖琴,葛继科,郑宏.**基于本体构建的协同推荐研究**[J]. 现代图书情报技术, 2008,24(9): 53-57
[13] 王强.**基于事务标识列表的关联规则挖掘算法**[J]. 现代图书情报技术, 2008,24(8): 63-69
[14] 夏立新,韩永青,张进.**基于本体的情报检索学科知识组织体系构建***[J]. 现代图书情报技术, 2008,24(12): 80-85
[15] 刘春艳,陈淑萍,伍玉成 .**基于SKOS的叙词表到本体的转换研究**[J]. 现代图书情报技术, 2007,2(5): 32-35