现代图书情报技术 » 2013, Vol. 29 » Issue (10) :27-30

文本分类中TF-IDF方法的改进研究

覃世安, 李法运

福州大学公共管理学院 福州 350108

Qin Shian, Li Fayun

School of Public Administration and Policy, Fuzhou University, Fuzhou 350108, China

- 摘要
- 参考文献
- 相关文章

Download: PDF (322KB) HTML (1KB) Export: BibTeX or EndNote (RIS) Supporting Info

摘要 针对TF-IDF在待分类文本类的数量分布不均时提取特征值效果差的问题,提出使用特征值在类间出现的概率比代替特征值在类间出现的次数比以改进TF-IDF算法。实验证明利用改进后的TF-IDF方法提取网页文本特征值,并配合简单累加求和的分类器,使得网页文本分类的准确率有明显提高,且分类速度加快。

关键词： 概率 TF-IDF 网页 文本分类

Abstract： When the count of one class is much more than another class's, the result of IDF in TF-IDF goes the wrong way according to its design idea. This paper solves the problem by using probability to change TF-IDF algorithm. In the end, the experiment proves that the solution mentioned above is good at classifying webpage text through a simple way to cumulative sum the value of characteristic words and the speed is faster and the accuracy rate is promoted.

Keywords： Probability, TF-IDF, Webpage, Text classification

收稿日期: 2013-06-17;

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
- 覃世安
- 李法运

引用本文:

覃世安, 李法运 .文本分类中TF-IDF方法的改进研究[J] 现代图书情报技术, 2013,V29(10): 27-30

Qin Shian, Li Fayun .Improved TF-IDF Method in Text Classification[J] , 2013,V29(10): 27-30

链接本文:

http://www.infotech.ac.cn/CN/ 或 http://www.infotech.ac.cn/CN/Y2013/V29/I10/27

[1] Sebastiani F. Machine Learning in Automated Text Categorization[J]. *ACM Computing Surveys (CSUR)*, 2002,34(1):1-47.

[2] 鲁松,李晓黎,白硕. 文档中词语权重计算方法的改进[J]. 中文信息学报, 2000,14(6): 8-13.(Lu Song,Li Xiaoli,Bai Shuo.An Improved Approach to Weighting Terms in Text[J].*Journal of Chinese Information Processing*, 2000,14(6): 8-13.)

[3] 罗欣, 夏德麟, 晏蒲柳. 基于词频差异的特征选取及改进的TF-IDF公式[J]. 计算机应用, 2005, 25(9): 2031-2033. (Luo Xin, Xia Delin,Yan Puliu. Improved Feature Selection Method and TF-IDF Formula Based on Word Frequency Differentia[J]. *Journal of Computer Applications*, 2005, 25(9): 2031-2033.)

[4] 张保富, 施化吉, 马素琴. 基于 TFIDF文本特征加权方法的改进研究[J]. 计算机应用与软件, 2011, 28(2):17-20.(Zhang Baofu,Shi Huaji,Ma Suqin. An Improved Text Feature Weighting Algorithm Based on TFIDF[J]. *Computer Applications and Software*, 2011,28(2): 17-20.)

[5] Forman G. BNS Feature Scaling: An Improved Representation over tf-idf for SVM Text Classification[C].In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. ACM, 2008: 263-270.

[6] Lan M, Tan C L, Low H B, et al. A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines[C].In: *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*. New York, NY, USA: ACM, 2005: 1032-1033.

[7] Oren N. Reexamining tf. idf Based Information Retrieval with Genetic Programming[C].In: *Proceedings of the 2002 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology*. Republic of South Africa: South African Institute for Computer Scientists and Information Technologists, 2002: 224-234.

[8] Aizawa A. An Information-theoretic Perspective of tf-idf Measures[J].*Information Processing and Management*,2003,39(1):45-65.

[9] 梁之舜, 邓集贤, 杨维权,等.概率论及数理统计[M].北京:高等教育出版社,1988.(Liang Zhishun, Deng Jixian, Yang Weiquan, et al. Probability Theory and Mathematical Statistics[M].Beijing:Higher Education Press,1988.)

[10] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9): 1848-1859.(Su Jinshu,Zhang Bofeng,Xu Xin. Advances in Machine Learning Based Text Categorization[J].*Journal of Software*,2006,17(9):1848-1859.)

[11] Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese Lexical Analyzer ICTCLAS[C].In: *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics,2003,17:184-187.

[12] 张玉芳,彭时名,吕佳. 基于文本分类TFIDF方法的改进与应用[J]. 计算机工程,2006,32(19): 76-78.(Zhang Yufang,Peng Shiming,Lv Jia. Improvement and Application of TFIDF Method Based on Text Classification[J].*Computer Engineering*,2006,32(19):76-78.)

[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] 路永和, 李焰锋.多因素影响的特征选择方法[J]. 现代图书情报技术, 2013,(5): 34-39

[3] 王昊, 李思舒, 邓三鸿.基于N-Gram的文本语种识别研究[J]. 现代图书情报技术, 2013,(4): 54-61

[4] 屈鹏, 王惠临.专利文本分类的基础问题研究[J]. 现代图书情报技术, 2013,29(3): 38-44

[5] 张倩, 刘怀亮.一种基于半监督学习的短文本分类方法[J]. 现代图书情报技术, 2013,29(2): 30-35

[6] 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

[7] 徐坤, 曹锦丹, 毕强.FCA在医学领域文本分类中的研究和应用[J]. 现代图书情报技术, 2012,28(3): 23-26

[8] 范云杰, 刘怀亮.基于维基百科的中文短文本分类研究[J]. 现代图书情报技术, 2012,28(3): 47-52

[9] 路永和, 何新宇.锐化高斯模板在文本特征项权重调整方法中的应用[J]. 现代图书情报技术, 2012,(12): 39-44

[10] 路永和, 曹利朝.基于粒子群优化的文本特征选择方法[J]. 现代图书情报技术, 2011,27(7/8): 76-81

[11] 徐健, 温浩胜.人才网页自动识别系统研究[J]. 现代图书情报技术, 2011,27(6): 20-26

[12] 谷俊, 王昊.基于领域中文文本的术语抽取方法研究[J]. 现代图书情报技术, 2011,27(4): 29-34

[13] 马芳.基于RBFNN的专利自动分类研究[J]. 现代图书情报技术, 2011,27(12): 58-63

[14] 胡泽文, 王效岳, 白如江.基于SUMO和WordNet本体集成的文本分类模型研究[J]. 现代图书情报技术, 2011,27(1): 31-38

[15] 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