P.O.Box 8718, Beijing 100080, China	Journal of Software, May 2005,16(5):1012-1020
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2005 by The Editorial Department of Journal of Software

DF还是IDF?主特征模型在Web信息检索中的使用

张 敏, 马少平, 宋睿华

张 敏1,2, 马少平1,2, 宋睿华1,2
1(清华大学 计算机科学与技术系,北京 100084)
2(清华大学 智能技术与系统国家重点实验室,北京 100084)
作者简介: 张敏(1977－),女,宁夏银川人,博士,讲师,主要研究领域为信息检索,机器学习;马少平(1961－),男,教授,博士生导师,主要研究领域为汉字识别,古籍数字化,信息检索;宋睿华(1978－),女,硕士生,主要研究领域为信息检索.
联系人: 张 敏 Phn: +86-10-62783191, E-mail: z-m@tsinghua.edu.cn, http://www.tsinghua.edu.cn

Received 2003-10-14; Accepted 2004-09-08

Abstract
In Web information retrieval (IR), input queries are too short and fuzzy to describe user request, which leads to the mismatch problem between user query and the documents full of redundancy and noise. This paper first studies the feature of web documents information and proposes the concepts of primary feature word, primary feature field and primary feature space (PFS). Then a new PFS query term weighting scheme is proposed, which takes document frequency (DF) into account instead of the traditional IDF factor. Finally, a combination strategy of term weighting is given. Using this PFS model, three groups of experiments have been performed on 10G and 19G large scale Web collections with TREC9, TREC10 and TREC11 standard tests of Web tracks. Comparative studies indicate that the new DF-related PFS term weighting improves the system performance consistently and effectively in terms of recall, top n precision and mean average precision. At most 18.6% improvement has been made.

Zhang M, Ma SP, Song RH. DF or IDF? On the use of primary feature model for Web information retrieval. *Journal of Software*, 2005,16(5):1012-1020.
DOI: 10.1360/jos161012
http://www.jos.org.cn/1000-9825/16/1012.htm

摘要
Web信息检索的难点之一就是简短、模糊的用户查询与存在大量冗余和噪声的文档之间的不匹配.对Web文档信息特征进行分析,提出Web文档主特征词、主特征域和主特征空间的概念,在该空间上使用文档频度DF(document frequency)信息而非传统意义上的IDF(inverse document frequency)信息进行权值计算,并给出一个改进的相似度计算模型.使用该模型在10G和19G的两个大规模Web文档集合上进行了3组标准测试.比较实验表明,与传统IDF思想相比,在各项评价指标上,DF相关的主特征权值计算方法都能始终较大幅度地提高系统性能,最大达到18.6%的性能改善.

基金项目: Supported by the National Natural Science Foundation of China under Grant Nos.60223004, 60321002, 60303005 (国家自然科学基金)

References:

[1] Anick PG. Adapting a full-text information retrieval system to computer the troubleshooting domain. In: Croft WB, van Rijsbergen CJ, eds. Proc. of the 17th Annual Int'l ACM-SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'94). ACM Press, 1994. 349-358.

[2] Croft WB, Cook R, Wilder D. Providing government information on the Internet: Experience with THOMAS. In: Proc. of the 2nd Int'l Conf. in Theory and Practice of Digital Libraries (DL'95). Texas, 1995. 19-24. http://csdl.tamu.edu/DL95/papers/croft/croft.html

[3] Stefan K, Armin H, Markus J, Andreas D. Improving document retrieval by automatic query expansion using collaborative learning of term-based concepts. Lecture Notes in Computer Science 2423, 2002. 376-387.

[4] Zhang M. Study on Web text information retrieval [Ph.D. Thesis]. Beijing: Tsinghua University, 2003 (in Chinese with English abstract).

[5] Moffat A, Davis R, Wilkinson R, Zobel J. Retrieval of partial documents. In: Harman D, ed. Proc. of the 2nd Text Retrieval Conf. (TREC 2). Gaithersburg: National Institute of Standards and Technology Special Publication, 1994. 181-191.

[6] Srinivasa S, Bhatt PCP. Introduction to Web information retrieval: A user perspective. Journal of Science Education, 2002,7(6): 27-38.

[7] Meng M, Yu C, Liu KL. Building efficient and effective metasearch engines. ACM Computing Surveys, 2002,34(1):48-89.

[8] Glover E, Tsioutsiouliklis K, Lawrence S, Pennock D, Flake G. Using Web structure for classifying and describing Web pages. In: Proc. of the Int'l World Wide Web Conf. (www 2002). Hawaii: ACM Press, 2002. 562-569. http://www2002.org/CDROM/refereed/504/index.html

[9] Cutler M, Shih Y, Meng W. Using the structure of HTML documents to improve retrieval. In: Proc. of the USENIX Symp. on Internet Technologies and Systems (NISTS'97). 1997. 241-251. http://www.usenix.org/publications/library/proceedings/usits97/full_papers/cutler/cutler.pdf

[10] Newby GB. Information space based on HTML structure. In: Vorhees E, ed. Proc. of the 9th Text Retrieval Conf. (TREC 9). Gaithersburg: National Institute of Standards and Technology Special Publication, 2000. 601-610.

[11] Ricardo BY, Berthier RN. Modern Information Retrieval. New York: Addison-Wesley, ACM Press, 1999. 19-34.

[12] Robertson SE, Walker S. Microsoft cambridge at TREC 9: Filtering track. In: Vorhees E, ed. Proc. of the 9th Text Retrieval Conf. (TREC 9). Gaithersburg: National Institute of Standards and Technology Special Publication, 2000. 25-33.

[13] Bailey P, Craswell N, Hawking D. Engineering a multi-purpose test collection for Web retrieval experiments. Information Proceeding and Management, 2003,39(6):853-871.

[14] Craswell N, Hawknig D. Overview of the TREC 2002 Web track. In: Vorhees E, ed. Proc. of the 11th Text Retrieval Conf. 2002 (TREC 2002). Gaithersburg: National Institute of Standards and Technology Special Publication, 2002. 61-68.

附中文参考文献:
[4] 张敏.Web文本信息检索方法研究[博士学位论文].北京:清华大学,2003.