P.O.Box 8718, Beijing 100080, China	Journal of Software May 2003,14(5):976-983
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2003 by The Editorial Department of Journal of Software

A Method to Query Document Database by Content and Structure

WANG Xiao-Ling, WEN Ji-Rong, LUAN Jin-Feng, MA Wei-Ying, DONG Yi-Sheng

Full-Text PDF Submission Back

WANG Xiao-Ling1, WEN Ji-Rong2, LUAN Jin-Feng1, MA Wei-Ying2, DONG Yi-Sheng1 1(Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China)2(Microsoft Research Asia, Beijing 100080, China)

Authors information: WANG Xiao-Ling was born in 1975. She is a Ph.D. candidate at the Department of Computer Science, Southeast University. Her current research interests include database theory and XML. WEN Ji-Rong is a researcher in Microsoft Research, China. His research interested areas are batabase theory and information retrieval. LUAN Jin-Feng was born in 1974. His research interested areas are artifical intelligence and communication. MA Wei-Ying is a researcher in Microsoft Research, China. His research interested areas are data mining and multimedia management theory. DONG Yi-Sheng was born in 1940. His current research interests include database theory and informaiton process.

Corresponding author: WANG Xiao-Ling, Phn: 86-25-6896725, E-mail: wxling@yahoo.com

Received 2002-04-04; Accepted 2002-10-17

Abstract

Structured documents are made up of a few logical components, such as title, sections, subsections and paragraphs. The components in each structured document can be represented by an ordered tree model, which can also be viewed as a hierarchical concept relationship. To meet the user's requirements for more precise and concentrated search results, the retrieval techniques should allow the user to retrieve document components with varying granularity. This paper presents a method to query document database by content and structure. The key idea is to construct a more comprehensive similarity function by taking advantage of the inherent hierarchical structure in documents. This work combines Information Retrieval techniques, semi-structured data query and proximate search for document documents. The proposed method is evaluated on the Encarta encyclopedia document set and the experimental results show that it can provide more accurate and focused answers than traditional document retrieval methods.

Wang XL, Wen JR, Luan JF, Ma WY, Dong YS. A method to query document database by content and structure. *Journal of Software*, 2003,14(5):976~983.

http://www.jos.org.cn/1000-9825/14/976.htm

摘要

文档是有一定逻辑结构的,标题、章节、段落等这些概念是文档的内在逻辑.不同的用户对文档的检索,有不同的需求,检索系统如何提供有意义的信息,一直是研究的中心任务.结合文档的结构和内容,对结构化文件的检索,提出了一种新的计算相似度的方法.这种方法可以提供多粒度的文档内容的检索,包括从单词、短语到段落或者章节.基于这种方法实现了一个问题回答系统,测试集是微软的百科全书Encarta,通过与传统方法实验比较,证明通过这种方法检索的文章片断更合理、更有效.

References:

[1] Extensible Markup Language (XML). http://www.w3c.org/XML/.

[2] Kaszkiel M, Zobel J, Sacks-Davis R. Efficient passage ranking for document databases. ACM Transactions on Information System, 1999,17(4):406~439.

[3] Clarke CLA, Cormack GV. Shortest-Substring retrieval and ranking. ACM Transactions on Information System, 2000,18(1):44~78.

[4] Cooper RJ, Rijger SM. A simple question answering system. In: Proceedings of the TREC-9. NIST Special Publication, 2000. http://www.doc.ic.ac.uk/~srueger/index.html.

[5] McHugh J, Widom J. Query optimization for XML. In: Proceedings of the 25th International Conference on Very large Data Bases. Edinburgh, Scotland, 1999. 315~326.

[6] Goldman R, McHugh J, Widom J. From semistructured data to XML: Migrating the lore data model and query language. In: Proceedings of the 2nd International Workshop on the Web and Databases (WebDB'99). Philadelphia, 1999. 25~30.

[7] XML query. http://www.w3c.org/XML/Query.

[8] Wang XL, Wen JR, Liu WY, Dong YS. Enhancive index for structured document retrieval. In: Proceedings of the12th International Workshop on Research Issues on Data Engineering: Engineering E-Commerce/E-Business Systems (RIDE-2EC 2002, Workshop of ICDE 02). San Jose, California: IEEE, 2002. 34~38.