P.O.Box 8718, Beijing 100080, China
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn

Journal of Software, April 2007,18(4):919-932
ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
Copyright © 2007 by *Journal of Software*

XML信息检索中最小子树根节点问题的分层算法

孔令波, 唐世渭, 杨冬青, 王腾蛟, 高 军

Full-Text PDF Submission Back

孔令波1, 唐世渭1,2, 杨冬青1, 王腾蛟1, 高 军1
1(北京大学 计算机科学技术系,北京 100871)
2(北京大学 视觉与听觉信息处理国家重点实验室,北京 100871)
作者简介: 孔令波(1974－),男,山东日照人,博士生,主要研究领域为关系数据库实现技术,XML数据处理技术,数据挖掘.唐世渭(1939－),男,教授,博士生导师,CCF高级会员,主要研究领域为数据库,半结构化数据,Web数据集成,数据挖掘.杨冬青(1945－),女,教授,博士生导师,CCF高级会员,主要研究领域为数据库,数据仓库,Web数据集成,移动数据挖掘.王腾蛟(1973－),男,博士,副教授,CCF高级会员,主要研究领域为数据库,数据仓库,Web数据集成,数据挖掘.高军(1975－),男,博士,副教授,主要研究领域为数据库,数据仓库,半结构化数据,Web数据集成,移动数据挖掘..
联系人: 孔令波 Phn: +86-10-62755440, E-mail: lbkong@db.pku.edu.cn

Received 2005-10-20; Accepted 2006-04-27

Abstract

SLCA (smallest lowest common ancestor) problem is a basic task of keyword search in XML information retrieval. It means to find all the nodes corresponding to the tightest subtrees in XML data, which involves the given keywords. Xu, et al., illustrate three algorithms-Indexed lookup eager (ILE), stack algorithm and scan eager (SE), and manifest that ILE has the best performance. Different from the complicated-B+-tree-based ILE algorithm, this paper proposes a layered solution for SLCA problem, named as LISA (layered intersection scan algorithm). It benefits from the distribution rule of SLCA nodes in XML tree, and calculates the SLCA nodes level by level (the deepest level runs first). That is, based on the retrieved Dewey codes corresponding to given keywords, the Dewey codes of SLCA nodes can be gotten by intersecting the prefix Dewey codes of each level. Compared with the ILE algorithm, LISA solutions need not sophisticated data structures, and have comparatively runtime performance. There are two instances following the LISA idea, called LISA I and LISA II respectively. They are distinguished from each other according to whether keeping Dewey codes in computation or transforming Dewey codes into integer sequences. Extensive experiments evaluate the performance of algorithms and prove the efficiency of LISA II.

Kong LB, Tang SW, Yang DQ, Wang TJ, Gao J. Layered solution for SLCA problem in XML information retrieval. *Journal of Software*, 2007,18(4):919-932.
DOI: 10.1360/jos180919
http://www.jos.org.cn/1000-9825/18/919.htm

摘要

最小子树根节点问题(smallest lowest common ancestor,简称SLCA)是实现XML信息检索研究中关键字查询的一个基本问题,其主旨就是求解所有包含给定关键字的紧致子树的根节点.XU等人给出了3种算法—基于索引的搜索算法(indexed lookup eager,简称ILE)、基于堆栈的算法以及基于扫描的算法(scan eager,简称SE),并通过实验证明ILE算法具有最好的表现.与基于B+树索引结构的ILE算法不同,所给出的新算法,称为LISA (layered intersection scan algorithm)方法.该方法基于SLCA节点按"层"分布的规律,采取了逐层求解SLCA节点的思路,即在获取了包含关键字的节点的Dewey码集合后,通过计算对应于不同关键字、不同层次的Dewey码前缀集合的交集,可以得到对应不同层的SLCA节点.与ILE相比,LISA除了只需对应于关键字的节点集合信息以外,不再需要其他复杂的辅助数据结构——全部的信息只是对应不同关键字的Dewey码集合以及排序操

作.同时,给出了两种实际的算法:LISA I和LISA II,二者的区别在于是否采用Dewey编码到整数的转换.其中,LISA II更具有满意的性能.

基金项目: Supported by the National Natural Science Foundation of China under Grant No.60503037 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2005AA4Z307 (国家高技术研究发展计划(863)); the Beijing Natural Science Foundation of China under Grant No.4062018 (北京市自然科学基金)

References:

[1] Meng XF, Zhou LX, Wang S. State of the art and trends in database research. Journal of Software, 2004,15(12):1822-1836 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/1822.htm

[2] Xu Y, Papakonstantinou Y. Efficient keyword search for smallest LCAs in XML databases. In: Ozcan F, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). Baltimore: ACM Press, 2005. 537-538.

[3] Tatarinov I, Viglas SD, Beyer K, Shanmugasundaram J, Shekita E, Zhang C. Storing and querying ordered xml using a relational database system. In: Franklin MJ, Moon B, Ailamaki A, eds. Proc. of the 2002 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). Madison: ACM Press, 2002. 204-215.

[4] Kong LB, Tang SW, Yang DQ, Wang TJ, Gao J. XML indices. Journal of Software, 2005,16(12):2063-2079 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/2063.htm

[5] Clark J, DeRose S. XML path language (XPath) version 1.0 w3c recommendation. World Wide Web Consortium, 1999. http://www.w3.org/TR/xpath

[6] Barg M, Wong RK. Structural proximity searching for large collections semi-structured data. In: Proc. of the ACM Conf. on Information and Knowledge Management (CIKM 2001). Atlanta: ACM Press, 2001. 175-182.

[7] Cohen S, Mamou J, Kanza Y, Sagiv Y. Xsearch: A semantic search engine for XML. In: Freytag JC, Lockemann PC, Abiteboul S, Carey MJ, Selinger PG, Heuer A, eds. Proc. of the 29th Int'l Conf. on Very Large Data Bases (VLDB). Berlin: Morgan Kaufmann Publishers, 2003. 45-56.

[8] Curtmola E, Amer-Yahia S, Brown P, Fernàndez M. GalaTex: A conformant implementation of the XQuery FullText language. In: Florescu D, Pirahesh H, eds. Informal Proc. of the 2nd Int'l Workshop on XQuery Implementation, Experience, and Perspectives (XIME-P). Baltimore: ACM Press, 2005. http://www.galaxquery.com/galatex/

[9] Amer-Yahia S, Botev C, Shanmugasundaram J. TeXQuery: A FullText search extension to XQuery. In: Feldman SI, Uretsky M, Najork M, Wills CE, eds. Proc. of the 13th Conf. on World Wide Web (WWW). New York: ACM Press, 2004. 583-594.

[10] Amer-Yahia S, Lakshmanan LV, Pandit S. FleXPath: Flexible structure and full-text querying for XML. In: Weikum G, K-nig AC, De-loch S, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). Paris: ACM Press, 2004. 83-94.

[11] Fuhr N, Gro-johann K. XIRQL: A query language for information retrieval in XML documents. In: Croft WB, Harper DJ, Kraft DH, Zobel J, eds. Proc. of the 24th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR). New Orleans: ACM Press, 2001. 172-180.

[12] Florescu D, Kossmann D, Manolescu I. Integrating keyword search into XML query processing. In: Albert V, et al., eds. Proc. of the 9th Int'l World Wide Web Conf. Amsterdam: ACM Press, 2000. 119-136.

[13] Balmin A, Papakonstantinou Y, Hristidis V. A system for keyword proximity search on XML databases. In: Freytag JC, Lockemann PC, Abiteboul S, Carey MJ, Selinger PG, Heuer A, eds. Proc. of the 29th Int'l Conf. on Very Large Data Bases (VLDB). Berlin: Morgan Kaufmann Publishers, 2003. 1069-1072.

[14] Weigel F, Meuss H, Schulz KU, Bry F. Content and structure in indexing and ranking XML. In: Amer-Yahia S, Gravano L, eds. Proc. of the 7th Int'l Workshop on the Web and Databases (WebDB). Paris: ACM Press, 2004. 67-72.

[15] Guo L, Shao F, Botev C, Shanmugasundaram J. XRANK: Ranked keyword search over XML documents. In: Halevy AY, Ives ZG, Doan A, eds. Proc. of the 2003 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). San Diego: ACM Press, 2003. 16-27.

[16] Botev C, Shanmugasundaram J. Context-Sensitive keyword search and ranking for XML. In: Doan AH, Neven F, McCann R, Bex GJ, eds. Proc. of the 8th Int'l Workshop on the Web & Databases (WebDB 2005). 2005. 115-120.

[17] Wang W, Jiang HF, Lu HJ, Yu JX. PBiTree coding and efficient processing of containment joins. In: Dayal U, Ramamritham K, Vijayaraman TM, eds. Proc. of the 19th Int'l Conf. on Data Engineering (ICDE). Bangalore: IEEE Computer Society, 2003. 391-402.

[18] Carmel D, Maarek YS, Mandelbrod M, Mass Y, Soffer A. Searching XML documents via XML fragments. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR). Toronto: ACM Press, 2003. 151-158.

[19] Amer-Yahia, S, Koudas N, Marian A, Srivastava D, Toman D. Structure and content scoring for XML. In: B-hm K, Jensen CS, Haas LM, Kersten ML, Larson P, Ooi BC, eds. Proc. of the 31st Int'l Conf. on Very Large Data Bases (VLDB). Trondheim: ACM Press, 2005. 361-372.

[20] Kailing K, Kriegel HP, Sch-nauer S, Seidl T. Efficient similarity search for hierarchical data in large databases. In: Bertino E, Christodoulakis S, Plexousakis D, et al., eds. Advances in Database Technology—EDBT 2004, the 9th Int'l Conf. on Extending Database Technology (EDBT). Heraklion: Springer-Verlag, 2004. 676-693.

[21] Yang R, Kalnis P, Tung AK. Similarity evaluation on tree-structured data. In: Ozcan F, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). Baltimore: ACM Press, 2005. 754-765.

[22] Augsten N, B-hlen MH, Gamper J. Approximate matching of hierarchical data using pq—Grams. In: B-hm K, Jensen CS, Haas LM, Kersten ML, Larson P, Ooi BC, eds. Proc. of the 31st Int'l Conf. on Very Large Data Bases (VLDB). Trondheim: ACM Press, 2005. 301-312.

附中文参考文献:
[1] 孟小峰,周龙骧,王珊.数据库技术发展趋.软件学报,2004,15(12):1822-1836. http://www.jos.org.cn/1000-9825/15/1822.htm

[4] 孔令波,唐世渭,杨冬青等.XML数据索引技术.软件学报,2005,16(12):2063-2079. http://www.jos.org.cn/1000-9825/16/2063.htm