P.O.Box 8718, Beijing 100080, China	Journal of Software, June 2007,18(6):
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2007 by *Journal of Software*

XML数据的查询技术

孔令波, 唐世渭, 杨冬青, 王腾蛟, 高 军

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,博士生,山东日照人,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,CCF高级会员,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@126.com, http://www.pku.edu.cn/lbkong

Received 2006-04-25; Accepted 2007-01-23

Abstract
XML has become the de facto standard for data representation and exchange for Web applications, such as digital library, Web service, and electronic business. How to retrieve interesting information from the promising XML data is an active research area. Among techniques in this area, the description of query patterns is a crucial section. This paper reviews the actualities of recent researches on this topic. It classifies the query descriptors into two categories, XML Query type and XML IR type (with three subcategories: XML IR/keyword, XML IR/fragment and XML IR/query), and concludes three popular problems: Twig pattern processing, SLCA (smallest lowest common ancestor) problem, and similarity measuring techniques for retrieved XML fragments. It analyzes the virtue and deficiency of related techniques based on their convenience for common users. And hereby it proposes four issues for further XML querying researches: structural keywords and corresponding structural similarity measuring, wiping off the redundancy in XML data processing between XML Query (including XML IR/query) and XML IR/keyword, theoretical discussion of XML Query and its realization, and the management of peculiar XML data.

Kong LB, Tang SW, Yang DQ, Wang TJ, Gao J. Querying techniques for XML data. *Journal of Software*, 2007,18(6):1400-1418.
DOI: 10.1360/jos181400
http://www.jos.org.cn/1000-9825/18/1400.htm

摘要
XML规范已成为当前网络应用(包括数字图书馆、Web服务以及电子商务)中事实上的数据表达、交换的标准.针对XML数据的查询在当前XML数据管理研究中占有重要的地位,也是当前XML数据处理研究领域的热点方向,相关的研究文献有很多.根据查询模式描述的不同,将当前XML查询技术归入两大类:XML Query方式和XML IR方式.后者又进而可分以为3个子类:XML IR/keyword方式、XML IR/fragment和XML IR/query方式,并从中挑选出3个研究者关注的问题进行了简述,它们是:Twig查询模式的处理、SLCA(smallest lowest common ancestor)节点的获取以及对所获取的XML片段相似性的度量.以方便普通用户使用为准则探讨了相关XML查询技术的优、缺点,将如下4个问题作为需要进一步关注的研究内容:结构化关键字查询及相应的结构相似性度量方法,如何消除XML Query查询处理模式(包含XML IR/query)和XML IR/keyword查询处理模式间

数据冗余的问题,XML Query查询方式的理论探讨及其实现以及针对特定应用的XML数据的有效管理.

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.2005AA4Z3070 (国家高技术研究发展计划(863)); the Beijing Natural Science Found 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] 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

[3] Abiteboul S, Quass D, McHugh J, Widom J, Wiener J. The Lorel query language for semistructured data. Int'l Journal on Digital Libraries, 1997,1(1):68-88.

[4] Deutsch A, Fernandez M, Florescu D, Levy A, Suciu D. A query language for XML. Computer Networks, 1999,31(11-16): 1155-1169.

[5] Ceri S, Comai S, Damiani E, Fraternali P, Paraboschi S, Tanca L. XML-GL: A graphical language for querying and restructuring XML documents. Computer Networks, 1999, 31(11-16):1171-1187.

[6] Chamberlin D, Robie J, Florescu D. Quilt: An XML query language for heterogeneous data sources. In: Suciu D, Vossen G, eds. Proc. of the Int'l Workshop on the Web and Databases (WebDB 2000). Dallas: Springer-Verlag, 2000. 1-25.

[7] Clark J, DeRose S. XML Path Language (XPath) Version 1.0 W3C Recommendation. World Wide Web Consortium, 1999. http://www.w3.org/TR/xpath

[8] Chamberlin D. XQuery: A query language for XML W3C working draft. Technical Report, WD-xquery-20010215, World Wide Web Consortium, 2001. http://www.w3.org/TR/xquery/

[9] Li QZ, Moon B. Indexing and querying XML data for regular path expressions. In: Apers PMG, Atzeni P, Ceri S, Paraboschi S, Ramamohanarao K, Snodgrass RT, eds. Proc. of the 27th Int'l Conf. on Very Large Data Bases (VLDB). Rome: Morgan Kaufmann Publishers, 2001. 361-370.

[10] Cooper BF, Sample1 N, Franklin MJ, Hjaltason GR, Shadmon M. A fast index for semistructured data. In: Apers PMG, Atzeni P, Ceri S, Paraboschi S, Ramamohanarao K, Snodgrass RT, eds. Proc. of the 27th Int'l Conf. on Very Large Data Bases (VLDB). Rome: Morgan Kaufmann Publishers, 2001. 341-350.

[11] Zhang C. Relational databases for XML indexing [Ph.D. Thesis]. Wisconsin: University of Wisconsin-Madison, 2002.

[12] Bruno N, Koudas N, Srivastava D. Holistic twig joins: Optimal XML pattern matching. 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. 310-321.

[13] Amer-Yahia S, Cho S, Lakshmanan LVS, Srivastava D. Minimization of tree pattern queries. In: Aref WG, ed. Proc. of the 2001 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). Santa Barbara: ACM Press, 2001. 497-508.

[14] Jiang HF, Wang W, Lu HJ, Yu JX. Holistic twig joins on Indexed XML documents. 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. 273-284.

[15] Chen ZY, Jagadish HV, Korn F, Koudas N. Counting twig matches in a tree. In: Young DC, ed. Proc. of the 17th Int'l Conf. on Data Engineering (ICDE). Heidelberg: IEEE Computer Society, 2001. 595-604.

[16] Lee DW, Srivastava D. Counting relaxed twig matches in a tree. In: Lee YJ, Li JZ, Whang KY, Lee DH, eds. Proc. of the 9th Int'l Conf. on Database Systems for Advances Applications (DASFAA). LNCS 2973, Springer-Verlag, 2004. 88-99.

[17] Jagadish HV, Al-Khalifa S. TIMBER: A native XML database. The VLDB Journal, 2002,11(4):274-291.

[18] Meng XF, Wang Y, Wang XF. Research on XML query optimization. Journal of Software, 2006,17(10):2069-2086 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/2069.htm

[19] Zhang N, -zsu MT, Ilyas IF, Aboulnaga A. FIX: Feature-Based indexing technique for XML documents. In: Dayal U, Whang KY, Lomet DB, et al., eds. Proc. of the 32nd Int'l Conf. on Very Large Data Bases (VLDB). Seoul: ACM Press, 2006. 259-270.

[20] Cho SR, Koudas N, Srivastava D. Meta-Data indexing for XPath location steps. In: Chaudhuri S, Hristidis V, Polyzotis N, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). Chicago: ACM Press, 2006. 455-466.

[21] Bird S, Chen Y, Davidson SB, Lee HJ, Zheng YF. Designing and evaluating an XPath dialect for linguistic queries. In: Liu L, Reuter A, Whang KY, et al., eds. Proc. of the 22nd Int'l Conf. on Data Engineering (ICDE). Atlanta: IEEE Computer Society, 2006. 52.

[22] Abiteboul S, Senellart P. Querying and updating probabilistic information in XML. In: Ioannidis YE, Scholl MH, eds. Advances in Database Technology, Proc. of the 10th Int'l Conf. on Extending Database Technology (EDBT 2006). Munich: Springer-Verlag, 2006. 1059-1068.

[23] Baeza-Yates R, Ribeiro-Neto B, et al. Modern Information Retrieval. Pearson Education Limited, 1999.

[24] 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.

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

[26] 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.

[27] Curtmola E, Amer-Yahia S, Brown P, Fernàndez M. GalaTex: A conformant implementation of the XQuery FullText language. In: Florescu D, Pirahesh H, eds. Proc. of the 2nd Int'l Workshop on XQuery Implementation, Experience, and Perspectives (XIME-P). Baltimore: ACM Press, 2005. 1024-1025.

[28] 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). Manhattan: ACM Press, 2004. 583-594.

[29] 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.

[30] 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.

[31] 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). Maison de la Chimie: ACM Press, 2004. 67-72.

[32] 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.

[33] 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.
[34] Florescu D, Kossmann D, Manolescu I. Integrating keyword search into XML query processing. http://www9.org/w9cdrom/index.html

[35] 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.

[36] Chinenyanga T, Kushmerick N. Expressive and efficient ranked querying of XML data. In: Mecca G, Siméon J, eds. Proc. of the 4th Int'l Workshop on the Web and Databases (WebDB 2001). Santa Barbara: ACM Press, 2001. 1-6.

[37] Theobald A, Weikum G. The index-based XXL search engine for querying XML data with relevance ranking. In: Jensen CS, Jeffery KG, Pokorny J, eds. Proc. of the 8th Conf. on Extending Database Technology (EDBT). Prague: Springer-Verlag, 2002. 477-495.

[38] Bremer JM, Gertz M. XQuery/IR: Integrating XML document and data retrieval. In: Fernandez MF, Papakonstantinou Y, eds. Proc. of the 5th Int'l Workshop on the Web and Databases (WebDB). Madison: ACM Press, 2002. 1-6.

[39] Hayashi Y, Tomita J, Kikui G. Searching text-rich XML documents with relevance ranking. In: Proc. of the SIGIR Workshop on XML and Information Retrieval. 2000. http://www.haifa.il.ibm.com/sigir00-xml/final-papers/Hayashi/hayashi.html

[40] Schmidt A, Kersten LM, Windhouwer M. Querying XML documents made easy: Nearest concept queries. In: Young DC, ed. Proc. of the 17th Int'l Conf. on Data Engineering (ICDE). Heidelberg: IEEE Computer Society, 2001. 595-604.

[41] Graupmann J, Schenkel R, Weikum G. The SphereSearch engine for unified ranked retrieval of heterogeneous XML and Web documents. In: B-hm K, Jensen CS, Haas LM, et al., eds. Proc. of the 31st Int'l Conf. on Very Large Data Bases (VLDB). Trondheim: ACM Press, 2005. 529-540.

[42] XQuery 1.0 and Xpath 2.0 fulltext. W3C Working Draft 1, 2006. http://www.w3.org/TR/xquery-full-text/

[43] Zhang C, Naughton J, DeWitt D, Luo Q, Lohman G. On supporting containment queries in relational database management systems. In: Aref WG, ed. Proc. of the 2001 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). Santa Barbara: ACM Press, 2001. 425-436.

[44] Wang J, Meng XF, Wang S. Structural join of XML based on range partitioning. Journal of Software, 2004,15(5):720-729 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/720.htm

[45] Wan CX, Liu YS, Xu SH, Liu XP, Lin DH. Indexing XML data based on region coding for efficient processing of structural joins. Chinese Journal of Computers, 2005,28(1):113-127 (in Chinese with English abstract). http://cjc.ict.ac.cn/qwjs/view.asp-id=1746

[46] Wang J, Meng XF, Wang Y, Wang S. Target node aimed path expression processing for XML data. Journal of Software, 2005, 16 (5):827-837 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/827.htm

[47] Lu JH, Ling TW, Chan CY, Chen T. From region encoding to extended Dewey: On efficient processing of XML twig pattern matching. 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. 193-204.

[48] Chen Y, Davidson SB, Zheng YF. BLAS: An efficient XPath processing system. 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. 47-58.

[49] Wang W, Wang HZ, Lu HJ, Jiang HF, Lin XM, Li JZ. Efficient processing of XML path queries using the disk-based F&B index. In: B-hm K, Jensen CS, Haas LM, et al., eds. Proc. of the 31st Int'l Conf. on Very Large Data Bases (VLDB). Trondheim: ACM Press, 2005. 145-156.

[50] Jiang HF, Lu HJ, Wang W. Efficient processing of XML twig queries with OR-predicates. 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. 59-70.

[51] Chen T, Lu JH, Ling TW. On boosting holism in XML twig pattern matching using structural indexing techniques. In: Ozcan F, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). Baltimore: ACM Press, 2005. 455-466.

[52] Shasha D, Zhang K. Approximate tree pattern matching. In: Apostolico A, Galil Z, ed. In: Proc. of the Pattern Matching Algorithms. Oxford University, 1997.
[53] Bille P. A survey on tree edit distance and related problems. Theoretical Computer Science, 2005,337(1-3):217-239.

[54] Joshi S, Agrawal N, Krishnapuram R, Negi S. A bag of paths model for measuring structural similarity in Web documents. In: Getoor L, Senator TE, Domingos P, et al., eds. Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD). Washington: ACM Press, 2003. 577-582.

[55] 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.

[56] Arvola P, Junkkari M, Kek-l-inen J. Generalized contextualization method for XML information retrieval. In: Herzog O, Schek H, Fuhr N, et al., eds. Proc. of the 2005 ACM CIKM Int'l Conf. on Information and Knowledge Management (CIKM). Bremen: ACM Press, 2005. 20-27.

[57] Wolff JE, Fl-rke H, Cremers AB. Searching and browsing collections of structural information. In: Proc. of the IEEE Advances in Digital Libraries (ADL 2000). Washington: ACM Press, 2000. 141-150.

[58] Guha S, Jagadish HV, Koudas N, Srivastava D, Yu T. Approximate XML joins. 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. 287-298.

[59] 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.

[60] 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.

[61] Schlieder T, Meuss H. Querying and ranking XML documents. Journal of the American Society for Information Science and Technology, 2002,53(6):489-503.

[62] Kailing K, Kriegel H, 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, Proc. of the 9th Int'l Conf. on Extending Database Technology (EDBT). Greece: Springer-Verlag, 2004. 676-693.

[63] Kotsakis E. Structured information retrieval in XML documents. In: Proc. of the 2002 ACM Symp. on Applied Computing (SAC). Madrid: ACM Press, 2002. 663-667.

[64] Wan CX, Liu YS. Efficient supporting XML query and keyword search in relational database systems. In: Meng XF, Su JW, Wang YJ, eds. Advances in Web-Age Information Management, Proc. of the 3rd Int'l Conf. (WAIM). LNCS 2419, Beijing: Springer-Verlag, 2002. 1-12.

[65] Hristidis V, Papakonstantinou Y, Balmin A. Keyword proximity search on XML graphs. In: Dayal U, Ramamritham K, Vijayaraman TM, eds. Proc. of the 19th Int'l Conf. on Data Engineering (ICDE). Bangalore: IEEE Computer Society, 2003. 367-378.

[66] Ré C, Siméon J, Fernández MF. A complete and efficient algebraic compiler for XQuery. In: Liu L, Reuter A, Whang KY, et al., eds. Proc. of the 22nd Int'l Conf. on Data Engineering (ICDE). Atlanta: IEEE Computer Society, 2006. 14.

[67] Zhang SH, Dyreson C. Symmetrically exploiting XML. In: Carr L, Roure DD, IyengarA, et al., eds. Proc. of the 15th Int'l Conf. on World Wide Web (WWW). Edinburgh: ACM Press, 2006. 103-111.

[68] Amer-Yahia S, Curtmola E, Deutsch A. Flexible and efficient XML search with complex full-text predicates. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). Chicago: ACM Press, 2006. 575-586.

附中文参考文献:

[1] 孟小峰,周龙骥,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

[2] 孔令波,唐世渭,杨冬青,王腾蛟,高军.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

[18] 孟小峰,王宇,王小锋.XML查询优化研究.软件学报,2006,17(10):2069-2086. http://www.jos.org.cn/1000-9825/17/2069.htm

[44] 王静,孟小峰,王珊.基于区域划分的XML结构连接.软件学报,2004,15(5):720-729. http://www.jos.org.cn/1000-9825/15/720.htm

[45] 万常选,刘云生,徐升华,刘喜平,林大海.基于区间编码的XML索引结构的有效结构连接.计算机学报,2005,28(1):113-127. http://cjc.ict.ac.cn/qwjs/view.asp-id=1746

[46] 王静,孟小峰,王宇,王珊.以目标节点为导向的XML路径查询处理.软件学报,2005,16(5):827-837. http://www.jos.org.cn/1000-9825/16/827.htm