P.O.Box 8718, Beijing 100080, China	Journal of Software, Oct. 2006,17(10):2069-2086
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2006 by *Journal of Software*

XML查询优化研究

孟小峰, 王 宇, 王小锋

孟小峰1, 王 宇2, 王小锋1
1(中国人民大学 信息学院,北京 100872)
2(河北大学 计算中心,保定 071002)
作者简介: 孟小峰(1964－),男,博士,教授,博士生导师,CCF高级会员,主要研究领域为Web数据集成,XML数据库,移动数据管理.王宇(1973－),女,博士,主要研究领域为Web数据管理,XML数据库.王小锋(1980－),女,硕士生,主要研究领域为XML数据库.
联系人: 孟小峰 Phn: +86-10-62519453, E-mail: xfmeng@ruc.edu.cn, http://www.ruc.edu.cn

Received 2006-01-19; Accepted 2006-04-17

Abstract
XML has become the de-facto standard for data representation and exchange on the World-Wide Web. Due to the nature of information on the Web and the inherent flexibility of XML, it is expected that much of the data encoded in XML will be semi-structured. Data on the internet is increasingly presented in XML format which enables researches on various kinds of XML storage model. Meanwhile, XML query optimization has become a hot research topic in database field. This paper gives an overview of the current status of technology for XML query optimization. The features of XML query optimization and key problems of research are also discussed deeply. Main aspects of current work on XML query optimization include XML algebra, cost model, complex path selectivity estimation, statistics information, and so on. Finally, this paper prospects future research directions and presents some viewpoints of XML query optimization.

Meng XF, Wang Y, Wang XF. Research on XML query optimization. *Journal of Software*, 2006,17(10):2069-2086.
DOI: 10.1360/jos172069
http://www.jos.org.cn/1000-9825/17/2069.htm

摘要
XML已经成为网络上信息描述和信息交换的标准.由于网络上信息的本质特性和XML数据内在的灵活性,很多用XML编码的数据都是半结构化的.随着XML应用得越来越广泛,人们提出了多种XML数据的存储模型.与此同时,XML的查询优化也是数据库领域研究的一个重要课题.综合论述了XML数据查询优化技术的现状,指出了XML查询优化的特点和研究的关键性问题.描述了查询优化技术各个方面的重要研究成果以及存在的问题,进一步展望了未来的研究方向,并在此基础上提出了对XML查询优化方法的一些观点.

基金项目：Supported by the National Natural Science Foundation of China under Grant Nos.60073014, 60273018 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.2003CB317000 (国家重点基础研究发展规划(973)); the Key Project of Chinese Ministry of Education under Grant No.03044 (国家教育部科学技术重点项目); the Program for New Century Excellent Talents in University (国家教育部新世纪优秀人才支持计划)

References:

[1] Beech D, Malhotra A, Rys M. A formal data model and algebra for XML. In: Beech D, Malhotra A, Rys M, eds. Note to the W3C XML Query Working Group. 1999. 1?26. http://www-db.stanford.edu/infoseminar/Archive/FallY99/malhotra-slides/malhotra.pdf

[2] Fernandez M, Simeon J, Suciu D, Wadler P. A data model and algebra for XML query. 1999. http://www.cs.bell-labs.com/wadler/topics/xml.html#algebra

[3] Kay M. XSL transformations (XSLT), Version 1.0. W3C Recommendation, 1999. http://www.w3.org/TR/xslt

[4] Fankhauser P, Fernandez M, Malhotra A, Rys M, Simeon J, Wadler P. XQuery 1.0 formal semantics. W3C Working Draft, 2002. http://www.w3.org/TR/query-semantics/

[5] Fernandez M, Robie J. XQuery 1.0 and XPath 2.0 data model. W3C Working Draft, 2002. http://www.w3.org/TR/query-datamodel/

[6] Mary FF, Jér?me S, Byron C, Amélie M, Gargi S. Implementing xquery 1.0: The galax experience. 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. Berlin: Morgan Kaufmann Publishers, 2003. 1077-1080.

[7] McHugh J, Abiteboul S, Goldman R, Quass D, Widom J. Lore: A database management system for semistructured data. SIGMOD Record, 1997,2(3):54-66.

[8] McHugh J, Widom J. Query optimization for XML. In: Atkinson MP, Orlowska ME, Valduriez P, Zdonik SB, Brodie ML, eds. Proc. of the 25th Int'l Conf. on Very Large Data Bases. Edinburgh: Morgan Kaufmann Publishers, 1999. 315-326.

[9] Jagadish VH, Al-Khalifa S, Lakshmanan L, Nierman A, Paparizos S, Patel J, Srivastava D, Wu YQ. Timber: A native XML database. The VLDB Journal, 2002,11(4):274-?291.

[10] Jagadish VH, Al-Khalifa S, Lakshmanan L, Srivastava D, Thompson K. Tax: A tree algebra for XML. In: Ghelli G, Grahne G, eds. Database Programming Languages, 8th Int'l Workshop, DBPL 2001. Frascati: Springer-Verlag, 2001. 149-164.

[11] Frasincar F, Houben GJ, Pau C. XAL: An algebra for XML query optimization. In: Zhou XF, ed. Proc. of the 13th Australasian Database Conf. (ADC 2002). Melbourne: Monash University, 2002.

[12] Zhang D, Dong YS. A data model and algebra for the Web. In: Proc. of the 10th Int'l Workshop on Database & Expert Systems Applications. Florence: IEEE Computer Society, 1999. 711-714.

[13] Liefke H. Horizontal query optimization on ordered semistructured data. In: Cluet S, Milo T, eds. Proc. of the ACM SIGMOD Workshop on The Web and Databases (WebDB'99). Philadelphia: ACM Press, 1999. 61-66.

[14] Mukhopadhyay P, Papakonstantinou Y. Mixing querying and navigation in MIX. In: Agrawal R, Dittrich K, Ngu AHH, eds. Proc. of the 18th Int'l Conf. on Data Engineering. San Jose: IEEE Computer Society, 2002. 245-254.

[15] Paparizos S, Al-Khalifa S, Jagadish HV, Nierman A, Wu YQ. A physical algebra for XML. Technical Report, University of Michigan, 2002.

[16] Christophides V, Cluet S, Moerkotte G. Evaluating queries with generalized path expressions. In: Jagadish HV, Mumick IS, eds. Proc. of the 1996 ACM SIGMOD Int'l Conf. on Management of Data. Montreal: ACM Press, 1996. 413-422.

[17] Buneman P, Fan W, Simen J, Weinstein S. Constraints for semistructured data and XML. ACM SIGMOD Record, 2001,30(1): 47-45.

[18] World Wide Web Consortium. XML path language (XPath) Version 1.0. W3C Recommendation, 1999. http://www.w3.org/TR/xpath.html

[19] Chamberlin D, Clark J, Florescu D, Robie J, Siméon J, Stefanescu M. XQuery 1.0: An XML query language. Technical Report, World Wide Web Consortium, W3C Working Draft, 2001.

[20] Deutsch A, Fernandez M, Florescu D, Levy A, Suciu D. A query language for XML. 2003. http://www.research.att.com/~mff/files/final.html

[21] Robie J, Lapp J, Schach D. XML query language (XQL). http://www.w3.org/TandS/QL/QL98/pp/xql.html

[22] Chamberlin D, Robie J, Florescu D. Quilt: An XML query language for heterogeneous data sources. In: Suciu D, Vossen G, eds. The World Wide Web and Databases, 3rd Int'l Workshop WebDB 2000. LNCS 1997, Springer-Verlag, 2001. 1-25.

[23] Li Q, 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. VLDB 2001, Proc. of the 27th Int'l Conf. on Very Large Data Bases. Roma: Morgan Kaufmann Publishers, 2001. 361-370.

[24] Chan C, Felber P, Garofalakis M, Rastogi R. Efficieng filtering of XML documents with Xpath expressions. In: Agrawal R, Dittrich K, Ngu AHH, eds. Proc. of the 18th Int'l Conf. on Data Engineering. San Jose: IEEE Computer Society, 2002. 235-244.

[25] Wood PT. On the equivalence of XML patterns. In: In: John W L, Verónica D, Ulrich F, Manfred K, Kung-K L, Catuscia P, Luís M P, Yehoshua S, Peter J S, eds. Proc. of the 1st Int'l Conf. on Computer on Computation Logic. LNAI 1861, Berlin: Springer-Verlag, 2000. 1152-1166.

[26] Florescu D, Levy AY Suciu D. Query containment for conjunctive queries with regular expressions. In: Popa L, ed. Proc. of the 17th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems. Seattle: ACM Press, 1998. 139-148.

[27] Calvanese D, Giacomo GD, Lenzerini M. On the decidability of query containment ander constraints. In: Popa L, ed. Proc. of the 17th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems. Seattle: ACM Press, 1998. 149-158.

[28] Wadler P. A formal semantics of patterns in XSLT. Markup Languages archive, 1999,2(2):183-202.

[29] Wood PT. Minimizing simple XPath expressions. 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. 13-18.

[30] Amer-Yahis S, Cho S, Lakshmanan LV, Srivastava D. Minimization of tree pattern queries. In: Aref WG, ed. Proc. of the SIGMOD 2001 Electronic. Santa Barbara: ACM Press, 2001. 497-508.

[31] Ramanan P. Efficient algorithms for minimizing tree pattern queries. In: Franklin MJ, Moon B, Ailamaki A, eds. Proc. of the 2002 ACM SIGMOD Int'l Conf. on Management of Data. Madison: ACM Press, 2002. 299-309.

[32] Flesca S, Furfaro F, Masciari E. On the minimization for Xpath queries. In: Freytag JC, Lockemann PC, Abiteboul S, Carey MJ, Selinger PG, Heuer A, eds. VLDB 2003, Proc. of the 29th Int'l Conf. on Very Large Data Bases. Berlin: Morgan Kaufmann Publishers, 2003. 153-164.

[33] Popa L, Deutsch A, Sahuguet A, Tannen V. A chase too far? In: Chen WD, Naughton JF, Bernstein PA, eds. Proc. of the 2000 ACM SIGMOD Int'l Conf. on Management of Data. Dallas: ACM Press, 2000. 273-284.

[34] Kwong A, Gertz M. Schema-Based optimization of XPath expressions. Technical Report, University of California, 2001.

[35] Lynch CA. Selectivity estimation and query optimization in large databases with highly skewed distributions of column values. In: Bancilhon F, DeWitt DJ, eds. 14th Int'l Conf. on Very Large Data Bases. Los Angeles: Morgan Kaufmann Publishers, 1988. 240-251.

[36] Haas PJ, Swami AN. Sequential sampling procedures for query size estimation. SIGMOD Record, 1992,21(2):341-350.

[37] Ling Y, Sun W. A supplement to sampling based methods for query size estimation in a database system. ACM SIGMOD Record, 1992,21(4):12-15.

[38] Muralikrishna M, DeWitt DJ. Equi-Depth histograms for estimating selectivity factors for multi-dimensional queries. SIGMOD Record, 1988,17(3):28-36.

[39] Zhang N, Hass PJ, Josifovski V, Lohman GM, Zhang C. Statistical learning techniques for costing XML queries. In: B?hm K, Jensen CS, Haas LM, Kersten ML, Larson PK, Ooi BC, eds. Proc. of the 31st Int'l Conf. on Very Large Data Bases. Trondheim: ACM Press, 2005. 289-300.

[40] Chen ZY, Jagadish HV, Korn F, Koudas N, Muthukrishnan S, Ng RT, Srivastava D. Counting twig matches in a tree. In: Young DC, ed. Proc. of the 17th Int'l Conf. on Data Engineering. Heidelberg: IEEE Computer Society, 2001. 595-604.

[41] Freire J, Haritsa JR, Ramanath M, Roy P, Siméon J. StatiX: Making XML count. In: Franklin MJ, Moon B, Ailamaki A, eds. Proc. of the 2002 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2002. 181-191.

[42] Polyzotis N, Garofalakis MN. Statistical synopses for graph-structured XML databases. In: Franklin MJ, Moon B, Ailamaki A, eds. Proc. of the 2002 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2002. 358-369.

[43] Goldman R, Widom J. DataGuides: Enabling query formulation and optimization in semistructured databases. In: Jarke M, Carey MJ, Dittrich KR, Lochovsky FH, Loucopoulos P, Jeusfeld MA, eds. VLDB'97, Proc. of the 23rd Int'l Conf. on Very Large Data Bases. Athens: Morgan Kaufmann Publishers, 1997. 436-445.

[44] Chen ZY, Korn F, Koudas N, Muthukrishnan S. Selectivity estimation for Boolean queries. In: Popa L, ed. Proc. of the 19th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems. Dallas: ACM Press, 2000. 216-225.

[45] Polyzotis N, Garofalakis M. Structure and value synopses for XML data graphs. In: Bressan S, Chaudhri AB, Lee ML, Yu JX, Lacroix Z, eds. Proc. of the 28th VLDB Conf. Hong Kong: Morgan Kaufmann Publishers, 2002. 466?-477.

[46] Polyzotis N, Garofalakis M, Iosnnidis Y. Selectivity estimation for XML twigs. In: Titsworth F, ed. Proc. of the 20th Int'l Conf. on Data Engineering, ICDE 2004. Boston: IEEE Computer Society, 2004. 264-275.

[47] Zhang N, Ozsu MT, Aboulnaga A, Ilyas IF. XSEED: Accurate and fast cardinality estimation for XPath queries. In: Ling L, Andreas R, Kyu-Y W, Jianjun Z, eds. Proc. of the 22nd Int'l Conf. on Data Engineering. Atlanta: IEEE Computer Society, 2006. 61.

[48] Schmidt AR, Waas F, Kersten ML, Florescu D, Manolescu I, Carey MJ, Busse R. The XML benchmark project. Technical Report, INS-R0103, CWI, 2001.

[49] Zhang C, Naughton JF, DeWitt DJ, Luo Q, Lohman GM. On supporting containment queries in relational database management systems. In: Aref WG, ed. Proc. of the 20th ACM SIGMOD Int'l Conf. on Management of Data. Santa Barbara: ACM Press, 2001. 425-436.

[50] Wu YQ, Patel JM, Jagadish HV. Estimating answer sizes for XML queries. In: Jensen CS, Jeffery KG, Pokorny J, Saltenis S, Bertino E, B?hm K, Jarke M, eds. Proc. of 8th Int'l Conf. on Extending Database Technology. Prague: Springer-Verlag, 2002. 590-608.

[51] Jiang HF, Lu HJ, Wang W, Yu JX. Containment join size estimation: Models and methods. In: Halevy AY, Ives ZG, Doan A, eds. Proc. of the 2003 ACM SIGMOD Int'l Conf. on Management of Data. San Diego: ACM Press, 2003. 145-156.

[52] Aboulnaga A, Alameldeen AR, Naughton JF. Estimating the selectivity of XML path expressions for Internet scale applications. 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. Roma: Morgan Kaufmann Publishers, 2001. 591-600.

[53] Matias Y, Vitter JC, Wang M. Wavelet-Based histograms for selectivity estimation. In: Haas LM, Tiwary A, eds. SIGMOD'98, Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Seattle: ACM Press, 1998. 448-459.

[54] Lim L, Wang M, Padmanabhan S, Vitter JS, Parr R. Xpath learner: An on-line self-tuning Markov histogram for XML path selectivity estimation. In: Bressan S, Chaudhri AB, Lee ML, Yu JX, Lacroix Z, eds. Proc. of the 28th Int'l Conf. on Very Large Data Bases. Hong Kong: Morgan Kaufmann Publishers, 2002. 442-453.

[55] Ramanath M, Zhang LZ, Freire J. Incremental maintence of schema-based XML statistics. In: Donald F. Shafer, eds. Proc. of the 21st IEEE Int'l Conf. on Data Engineering. Tokyo: IEEE Computer Society, 2005. 273-284.