P.O.Box 8718, Beijing 100080, China	Journal of Software, April. 2005,16(4):561-569
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

XML数据查询中值匹配查询代价估计算法

曲卫民, 孙 乐, 孙玉芳

曲卫民, 孙 乐, 孙玉芳
(中国科学院 软件研究所 系统软件与中文信息中心,北京 100080)
作者简介: 曲卫民(1977－),男,山西运城人,博士,主要研究领域为XML数据库,信息检索;孙乐(1971－),男,博士,副研究员,主要研究领域为信息检索,中文信息处理;孙玉芳(1947－2005),男,博士生导师,研究员,主要研究领域为系统软件,数据库理论.
联系人: 曲卫民 E-mail: quweimin@chinalife.com.cn, http://www.sonata.iscas.ac.cn

Received 2004-04-12; Accepted 2004-09-08

Abstract
Result size estimation of value predication in XML query is a multiple attributes dependent problem. It is different from the counterpart in relational database, for the multiple attributes in XML involve not only the value data, but also the structural information. To solve the problem, this paper proposes a wavelet-based histogram for the result size estimation of value predication in XML query. It also gives the way to identify the multi-dimensional dependent element set, to rewrite the value predication and value denotation of structural information. Experimental results show that the algorithm achieves on accurate result size estimation for value predication in XML query.

Qu WM, Sun L, Sun YF. A result size estimation algorithm for value predication in XML query. *Journal of Software*, 2005,16(4):561-569.
http://www.jos.org.cn/1000-9825/16/561.htm

摘要
XML数据查询中值匹配查询条件的查询代价估计问题是一种典型的多元素查询条件代价估计问题.它与传统关系型数据库中的多元素查询条件不同,因为XML数据中的值信息分布不仅与其他值信息分布相关,还与XML数据中的结构信息相关,而且当XML数据结构比较复杂时,可能会形成高维元素相关.针对以上问题,提出了一种面向XML数据的基于小波的多维直方图查询代价估计算法,并提出了确定XML数据中以某值元素为主键的相互依赖元组的方法,将值匹配条件改写为多元素查询条件的方法以及结构信息的值化方法.实验结果证明,提出的方法取得了较准确的查询代价估计结果.

基金项目: Supported by the National Natural Science Foundation of China under Grant No.60203007 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2003AA1Z2110 (国家高技术研究发展计划(863)); the New Star Plan of Science & Technology of Beijing of China under Grant No.H020820790130 (北京市科技新星计划)

References:

[1] McHugh J, Widom J. Query optimization for XML. In: Proc. of the VLDB. 1999. 315-326. http://www.vldb.org/conf/1999/P32.pdf

[2] Chen Z, Jagadish HV, Korn F, Koudas N, Muthukrishnan S, Ng RT, Srivastava D. Counting twig matches in a tree. In: Proc. of the ICDE. 2001. 595-604. http://citeseer.ist.psu.edu/chen01counting.html

[3] Aboulnaga A, Alameldeen AR, Estimating JN. The selectivity of XML path expressions for Internet scale applications. In: Proc. of the VLDB. 2001. 591-600.

[4] Wu Y, Patel JM, Jagadish HV. Estimating answer sizes for xml queries. In: Proc. of the EDBT. 2002.

[5] Polyzotis N, Garofalakis M. Statistical synopses for graph structured XML databases. In: Proc. of the SIGMOD. 2002.

[6] Freire J, Haritsa JR, Ramanath M, Roy P, Sim_eon J. StatiX: Making XML count. In: Proc. of the 2002 ACM SIGMOD Int'l. Conf. on Management of Data. 2002.

[7] Muralikrishna M, Dewitt DJ. Equi-Depth histograms for estimating selectivity factors for multi-dimensional queries. In: Proc. of the ACM SIGMOD Conf. 1988. 28-36.

[8] Poosala V, Ioannidis Y. Selectivity estimation without the attribute value independence assumption. Technical Report, Bell Labs, 1997.

[9] Deshpande A, Garofalakis M, Rastogi R. Independence is good: Dependency-Based histogram synopses for high-dimensional data. In: Proc. of the ACM SIGMOD 2001. 2001. 199-210. http://www.bell-labs.com/user/minos/Papers/sigmod01dbhist-cam.pdf

[10] Vitter JS, Wang M. Approximate computation of multidimensional aggregates of sparse data using wavelets. ACM SIGMOD, 1999.

[11] DBLP Computer Science Bibliography. http://www.informatik.uni-trier.de/~ley/db/

[12] Schmidt A, Waas F, Kersten M, Florescu D, Manolescu I, Carey M, Busse R. The XML benchmark project. Technical Report INSR0103, 2001.