基于树自动机的XPath在XML数据流上的高效执行

高 军, 杨冬青, 唐世渭, 王腾蛟

高 军, 杨冬青, 唐世渭, 王腾蛟

(北京大学 信息科学技术学院,北京 100871)

作者简介: 高军(1975－),男,山东临沂人,博士,讲师,主要研究领域为数据库与信息系统;杨冬青(1945－),女,教授,博士生导师,主要研究领域为数据库与信息系统;唐世渭(1939－),男,教授,博士生导师,主要研究领域为数据库与信息系统;王腾蛟(1973－),男,博士,讲师,主要研究领域为数据库与信息系统.

联系人: 高 军 Phn: +86-10-62765825, Fax: +86-10-62765822, E-mail: gaojun@db.pku.edu.cn, http://www.pku.edu.cn

Abstract

How to efficiently evaluate massive XPaths set over an XML stream is a fundamental problem in applications of the data stream. The current methods can not fully support the commonly used features of XPath, or can not meet the space and time requirement of the data stream applications. In this paper, a new tree automata based machine, XEBT, is proposed to solve the problem. Different from traditional ones, XEBT has the following features: First, it is based on tree automata with a powerful expressiveness, which can support Xpath {[]} without extra states or intermediate results; Second, XEBT supports many optimization strategies, including DTD based XPath tree automata construction, partial determination to reduce the concurrent states at running time with limited extra space costs, and the combination of bottom-up and top-down evaluation. Experimental results show that XEBT supports the complex Xpath and outperforms the former work in both efficiency and space cost.

摘要

如何在XML数据流上高效地执行大量的XPath查询成为数据流应用中一个迫切需要解决的关键问题.目前提出的算法或者不能完全支持XPath的常规特性,或者在算法的执行效率和空间代价上不能满足数据流应用的要求.提出了基于树自动机的XEBT机来解决这个问题.与传统方法相比,XEBT机具备如下特征:首先,XEBT机基于表达能力丰富的树自动机,无须附加中间状态,或保存中间结果,就能处理支持{[]}操作符的XPath;其次,XEBT机支持多种优化策略,包括基于DTD的XPath查询自动机的构造;在空间代价有限增加的情况下采用局部确定化减少并发执行的状态;采用自上而下和自下而上相结合的查询处理策略.实验结果表明,提出的方法能够支持复杂的XPath查询,在执行效率和空间代价方面优于传统算法.

References:

[1] Diao Y, Fischer P. YFilter: Efficient and scalable filtering of XML documents. In: Proc. of the 18th Int'l Conf. on Data Engineering. 2002. 341-345.

[2] Chan C, Felber P, Garofalakis M, Rastogi R. Efficient filtering of XML document with XPath expressions. In: Proc. of the Int'l Conf. on Data Engineering. San Jose: IEEE Computer Society, 2002. 235-244.

[3] Green TJ, Miklau G, Onizuka M, Suciu D. Processing XML streaming with deterministic automata. In: Calvanese D, Lenzerini M, Motwani R, eds. Proc. of the Int'l Conf. on Data Theory. LNCS 2572, Springer-Verlag, 2003. 173-189.

[4] Gupta AK, Suciu D. Stream processing of XPath queries with predicates. In: Halevy AY, Ives ZG, Doan AH, eds. Proc. of the 2003 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2003. 419-430.

[5] Nguyen B, Abiteboul S, Cobena G, Preda M. Monitoring XML data on the Web. In: Aref WG, ed. Proc. of the ACM/SIGMOD Conf. on Management of Data. 2001. 437-448.

[6] Chen J, Dewitt D, Tian F, Wang Y. NiagaraCQ: A scalable continuous query system for internet databases. In: Chen WD, Naughton JF, Bernstein PA, eds. Proc. of the ACM/SIGMOD Conf. Management of Data. ACM, 2000. 379-390.

[7] Clark J. XML Path language (XPath). 1999. Available from the W3C, http://www.w3.org/TR/XPath

[8] Neven F. Automata, logic, and XML. In: Proc. of the 16th Int'l Workshop Computer Science Logic. CSL, 2002. 2-26.

[9] Milo T, Suciu D, Vianu V. Typechecking for XML Transformers. In: Proc. of the PODS 2000. ACM, 2000. 11-22.

[10] Miklau G, Suciu D. Containment and equivalence for an XPath fragment. In: Popa L, ed. Proc. of the 21 Symp. on Principle of Database Systems. ACM, 2002. 65-76.

[11] Gao J, Yang DQ, Tang SW, Wang TJ. DTD based deterministic XPath rewriting and logical optimization. Journal of Software, 2004,15 (12):1860-1868 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/1860.htm

[12] NASA's Astronomical Data Center. ADC XML Resource Page. http://xml.gsfc.nasa.gov

附中文参考文献:
[11]   高军,杨冬青,唐世渭,王腾蛟.一种基于DTD的XPath逻辑优化方法.软件学报,2004,15(12):1860-1868.   http://www.jos.org.cn/1000-9825/15/1860.htm