P.O.Box 8718, Beijing 100080, China	Journal of Software, May 2005,16(5):869-877
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

面向XPath执行的XML数据流压缩方法

王腾蛟, 高 军, 杨冬青, 唐世渭, 刘云峰

王腾蛟, 高 军, 杨冬青, 唐世渭, 刘云峰
(北京大学 信息科学技术学院,北京 100871)
作者简介: 王腾蛟(1973－),男,山东济南人,博士,副教授,主要研究领域为数据库与信息系统;高军(1975－),男,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;刘云峰(1973－),女,博士生,主要研究领域为数据库与信息系统.
联系人: 王腾蛟 Phn: +86-10-62765823, E-mail: tjwang@pku.edu.cn, http://www.pku.edu.cn

Received 2003-10-13; Accepted 2004-03-02

Abstract
Because XML (extensible markup language) is self-described, there is much redundant structural information in XML data stream. How to compress XML data so as to reduce the network transfer cost and support XPath evaluation on the compressed data is a new area of research. The existing methods on XML compression require the multi-pass scan on data or can not support real time query processing on compressed data. In this paper, a novel compression method XSC (XML stream compression) is proposed to compress and decompress XML stream in real time. XSC constructs XML element event sequence dictionary and outputs the related index dynamically. When DTD is available, XSC can generate the XML element event sequence graph for producing more reasonable encoding before XML data stream is processed. The compressed XML data stream can be decomposed directly for XPath evaluation. Experimental results show that XSC outperforms other methods in compression ratio and compression efficiency, and the cost of XPath evaluation on compressed data stream is acceptable.

Wang TJ, Gao J, Yang DQ, Tang SW, Liu YF. XPath evaluation oriented XML data stream compression. Journal of Software, 2005,16(5):869-877.
DOI: 10.1360/jos160869
http://www.jos.org.cn/1000-9825/16/869.htm

摘要
由于XML(extensible markup language)本身是自描述的,所以XML数据流中存在大量冗余的结构信息.如何压缩XML数据流,使得在减少网络传输代价的同时有效支持压缩数据流上的查询处理,成为一个新的研究领域.目前已有的XML数据压缩技术,都需要扫描数据多遍,或者不支持数据流之上的实时查询处理.提出了一种XML数据流的压缩技术XSC(XML stream compression),实时完成XML数据流的压缩和解压缩,XSC动态构建XML元素事件序列字典并输出相关索引,能够根据XML数据流所遵从的DTD,产生XML元素事件序列图,在压缩扫描之前,产生更加合理的结构序列编码.压缩的XML数据流能够直接解压缩用于XPath的执行.实验表明,在XML数据流环境中,XSC在数据压缩率和压缩时间上要优于传统算法.

同时,在压缩数据之上查询的执行代价是可以接受的.

基金项目: Supported by the National High-Tech Research and Development Plan of China under Grant No.2002AA4Z3440 (国家高技术研究发展计划(863)); the National Grand Fundamental Research 973 Program of China under Grant No.G1999032705 (国家重点基础研究发展规划(973))

References:

[1] Hartmut L, Dan S. XMill: An efficient compressor for XML data. In: Weidong C, Jeffrey F, eds. Proc. of the SIGMOD 2000. Texas: ACM Press, 2000. 153-164.

[2] Pankaj MT, Jayant RH. XGRIND: A query friendly XML compressor. In: Proc. of the ICDE 2002. San Jose: IEEE Computer Society, 2002. 225-234.

[3] Jun KM, Myung JP, Chin WC. XPRESS: A queriable compression for XML data. In: Alon Y, Zachary G, eds. Proc. of the SIGMOD 2003. San Diego: ACM Press, 2003. 122-133.

[4] Jacob Z, Abraham L. A universal algorithm for sequence data compression. IEEE Trans. on Information Theory, 1977,23 (3):337-343.

[5] Jeffery SV. Design and analysis of dynamic Huffman codes. Journal of the ACM, 1987,34(4):825-845.

[6] Jean LG. GZIP. 2003. HTTP://www.gzip.com

[7] SwissProt Data Set. 1998. http://www.cs.washington.edu/research/xmldatasets/data/SwissProt/SwissProt.xml

[8] NASA Data Set. 2001. http://www.cs.washington.edu/research/xmldatasets/data/nasa/nasa.xml

[9] Tree Bank Data Set. 2002. http://www.cs.washington.edu/research/xmldatasets/data/treebank/treebank_e.xml

[10] Angel LD, Douglas L. XML generator. 1999. http://www.alphaworks.ibm.com/tech/xmlgenerator