Chapter 3
XML Compression

Chin-Wan Chung
Korea Advanced Institute of Science and Technology (KAIST,) Republic of Korea

Myung-Jae Park
Korea Advanced Institute of Science and Technology (KAIST,) Republic of Korea

Jihyun Lee
Korea Advanced Institute of Science and Technology (KAIST,) Republic of Korea

ABSTRACT

To effectively reduce the redundancy and verbosity of XML data, various studies for XML compression have been conducted. Especially, XML data management systems and applications require the support of direct query processing and update on compressed XML data, the stream based compression/decompression, and the reduction of the size of the compressed data. In order to fully support the various aspects of XML compression, existing XML compression techniques should be carefully examined and the additional requirements for XML compression techniques should be considered. In this chapter, the authors first classify existing representative XML compression techniques according to their characteristics. Second, they explain the details of XML specific compression techniques. Third, they summarize the performance of the compression techniques in terms of the compression ratio and the compression and decompression time. Lastly, they present some future research directions.

INTRODUCTION

XML (eXtensible Markup Language) (Bray et al., 2008) is a standardized markup language to represent and exchange data on the Web. Due to the proliferation of the Web, the usage of XML has tremendously increased. As a result, more data are produced as XML data and the size of XML data also increases. In order to efficiently manage such XML data, various types of researches on issues such as XML indexing, XML query processing, and XML storage have been conducted. Moreover, to effectively manage large sized XML data, the compression on XML data was required.

Data compression is the process of encoding data with a size smaller than that of the original data using specific encoding methods. As a result, data compression provides several important advantages. First, the storage size for compressed data is reduced compared with that for original data.

DOI: 10.4018/978-1-61520-727-5.ch003

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Second, the network bandwidth can be saved with the reduced data size since much more data can be transferred through the network within a given period of time. Lastly, the performance of query processing can be improved since the memory is efficiently utilized and the required number of disk I/Os is reduced.

According to those advantages of the data compression, the compression of XML data has been interested in various areas such as archiving, query processing, data dissemination and so on. Since an XML document is generally a text file, general text compression techniques such as gzip (Gailly & Adler, 2007) and bzip2 (Seward, 2008) can be employed to compress the XML document. However, an XML document is distinguished from a general plain text by the existence of the semantic structure in the XML document. Thus, various researches for XML specific compression have been conducted to effectively solve the redundancy and verbosity problems of XML data.

The XML specific compressors take advantage of the structure-awareness to improve the performance of the compression. In general, the structure of an XML document, which can be modeled as a tree composed of redundant elements and attributes, has a high regularity such that similar sub-trees repeatedly appear throughout the document. Also, data values enclosed by the same element have the same data type or are similar to each other. The regularity of data increases the compression ratio. Therefore, the XML specific techniques compress the syntactically or semantically partitioned structure and content using different encoding models to sufficiently take advantage of the local homogeneity. For the XML compression, the high compression ratio is an important goal of the XML compression. In addition, since XML data can be frequently queried and updated, the direct evaluation of queries and the direct update on compressed XML data should be possible. Since XML data are frequently exchanged in the Internet, the stream based compression/decompression should also be considered. Until now, most studies on the XML compression have focused on the achievement of high compression ratio and the direct query evaluation on compressed XML data. In order to fully support the above aspects of XML compression, existing XML compression techniques should be carefully examined and the additional requirements for XML compression techniques should be addressed.

The objective of this chapter is to provide a better understanding on relevant theoretical frameworks and an up-to-date research trend of the XML compression. To achieve this goal, this chapter contains the following contents.

First, various existing XML compression techniques are classified according to their characteristics. The classified categories are schema-dependent compression and schema-independent compression, non-queriable compression and queriable compression, and homomorphic compression and non-homomorphic compression. The characteristics of each category are introduced in detail.

Second, representative XML specific compression techniques including the latest ones such as XMill (Liefke & Suciu, 2000), XMLPPM (Cheney, 2001), XAUST (Hariharan & Shankar, 2006), XGRIND (Tolani & Haritsa, 2002), XQzip (Ng & Cheng, 2004), XPRESS (Min et al., 2003), XBzipIndex (Ferragina et al., 2006), XCQ (Ng et al., 2006b), XQueC (Arion et al., 2007), and ISX (Wong et al., 2007) are presented. Also, the compression performance of those compression techniques is summarized in terms of compression ratio, compression time, and decompression time. Furthermore, based on their characteristics and experimental evaluation, appropriate XML compression techniques for different environments are recommended.

Lastly, several applications which adapt the XML compression techniques are introduced. Also, based on the up-to-date research trend of XML compression techniques, important future research directions required for the development

17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/xml-compression/41499

Related Content

A Secure and Dynamic Mobile Identity Wallet Authorization Architecture Based on a XMPP Messaging Infrastructure
Alexandre B. Augustoand Manuel E. Correia (2013). *Innovations in XML Applications and Metadata Management: Advancing Technologies (pp. 21-37).*
www.irma-international.org/chapter/secure-dynamic-mobile-identity-wallet/73171

XML Benchmarking: The State of the Art and Possible Enhancements
Irena Mlynkova (2009). *Open and Novel Issues in XML Database Applications: Future Directions and Advanced Technologies (pp. 309-327).*
www.irma-international.org/chapter/xml-benchmarking-state-art-possible/27787

Temporal OCL: Meeting Specification Demands for Business Components
Stefan Conradand Klaus Turowski (2001). *Unified Modeling Language: Systems Analysis, Design and Development Issues (pp. 152-167).*
www.irma-international.org/chapter/temporal-ocl-meeting-specification-demands/30577

Formalizing and Analyzing UML Use Case Hierarchical Predicate Transition Nets
Xudong He (2005). *Advances in UML and XML-Based Software Evolution (pp. 154-183).*
www.irma-international.org/chapter/formalizing-analyzing-uml-use-case/4935

Towards Massive RDF Storage in NoSQL Databases: A Survey
Zongmin Maand Li Yan (2019). *Emerging Technologies and Applications in Data Processing and Management (pp. 263-284).*
www.irma-international.org/chapter/towards-massive-rdf-storage-in-nosql-databases/230693