Structure– and Content–Based Retrieval for XML Documents

Jae-Woo Chang
Chonbuk National University, South Korea

INTRODUCTION

The XML was proposed as a standard markup language to make Web documents in 1996 (Extensible Markup Language, 2000). It has as good an expressive power as SGML and is easy to use like HTML. Recently, it has been common for users to acquire through the Web a variety of multimedia documents written by XML. Meanwhile, because the number of XML documents is dramatically increasing, it is difficult to reach a specific XML document required by users. Moreover, an XML document not only has a logical and hierarchical structure in common, but also contains its multimedia data, such as image and video. Thus, it is necessary to retrieve XML documents based on both document structure and image content. For supporting the structure-based retrieval, it is necessary to design four efficient index structures, that is, keyword, structure, element, and attribute index, by indexing XML documents using a basic element unit. For supporting the content-based retrieval, it is necessary to design a high-dimensional index structure so as to store and retrieve both color and shape feature vectors efficiently.

BACKGROUND

Because an element is a basic unit that constitutes a structured (i.e., SGML or XML) document, it is essential to support not only retrieval based on element units but also retrieval based on logical inclusion relationships among elements. First, RMIT in Australia proposed a subtree model that indexes all the elements in a document and stores all the terms that appear in the elements (Lowe, Zobel & Sacks-Davis, 1995) so as to support five query types for structure-based retrieval in SGML documents. Secondly, SERI in South Korea proposed a K-ary Complete Tree Structure, which represents a SGML document as a K-ary complete tree (Han, Son, Chang & Zhao, 1999). In this method, a relationship between two elements can be acquired by calculation because each element corresponds to a node in a K-ary tree. Thirdly, University of Wisconsin in Madison proposed a new technique to use the position and depth of a tree node for indexing each occurrence of XML elements (Zhang, Naughton, Dewitt, Luo & Lohman, 2001). For this, the inverted index was used to enable ancestor queries to be answered in constant time. Fourthly, IBM T.J. Watson research center in Hawthorne proposed ViST, a novel index structure for searching XML documents (Wang, Park, Fan & Yu, 2003). The ViST made use of tree structures as the basic unit of query to avoid expensive join operations and provided a unified index on both text content and structure of XML documents. However, these four indexing techniques were supposed to handle tree data. Finally, University of Singapore proposed D(k)-Index, a structural summary for general graph structured documents (Chen, Lim & Ong, 2003). The D(k) index possesses the adaptive ability to adjust its structure according to the current query load, thus facilitating efficient update algorithms.

There have been a lot of studies on content-based retrieval techniques for multimedia or XML documents. First, the QBIC (Query By Image Content) project of IBM Almaden research center studied content-based image retrieval on a large online multimedia database (Flickner et al., 1995). The study supported various query types based on the visual image features such as color, texture, and shape. Secondly, the VisualSEEK project of Colombia University in the USA developed a system for content-based retrieval and browsing (Smith & Chang, 1996). Its purpose was an implementation of CBVQ (Content-Based Visual Query) that combines spatial locations of image objects and their colors. Thirdly, the Pennsylvania State University presented a comprehensive survey on the use of pattern recognition methods for content-based retrieval on image and video information (Antani, Kasturi & Jain, 2002). Finally, the Chonbuk National University in South Korea developed an XML document retrieval system that can support a unified retrieval based on both image content and document structure (Chang, 2002). Finally, the Chinese University of Hong Kong presented a multi-lingual digital video content management system, called iVIEW, for intelligent searching and access of English and Chinese video contents (Lyu, Yau & Sze, 2002). The iVIEW system allows full content indexing and retrieval of multi-lingual text, audio and video materials in XML documents.
Related Content

Collaborative Decision Making in Web-Based GIS
www.igi-global.com/chapter/collaborative-decision-making-web-based/4616?camid=4v1a

An Overview of Asynchronous Online Learning
www.igi-global.com/chapter/overview-asynchronous-online-learning/14009?camid=4v1a

Utility Based Frequent Pattern Extraction from Mobile Web Services Sequence
www.igi-global.com/article/utility-based-frequent-pattern-extraction-from-mobile-web-services-sequence/203007?camid=4v1a

A Multi-Agent Optimization Method for Preemptive Resource-Constrained Project Scheduling
Yongyi Shou, Wenjin Hu, Changtao Lai and Ying Ying (2019). International Journal of Information Technology Project Management (pp. 21-33).
www.igi-global.com/article/a-multi-agent-optimization-method-for-preemptive-resource-constrained-project-scheduling/215012?camid=4v1a