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 & Zhoo, 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 Univesity presented a comprehensive survey on the use of pattern recognition methods for content-based retrieval on image and video information (Antani, Kasturi & Jain, 2002). Fourthly, 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.

STRUCTURE- AND CONTENT-BASED RETRIEVAL

A structure- and content-based XML document retrieval system consists of five main parts: a preprocessing part for parsing XML documents and doing image segmentation, an indexing part for generating index keys of XML documents, a storage manager part for storing index information into a specific database, a unified retrieval part for finding results and integrating them into an unified one, and a user interface part for answering user queries by using a Web browser.

When XML documents are given, they are parsed, and image segmentation is done through the preprocessing part. The parsed document information is transported into the structure-based indexer in order to index its document structure consisting of element units. By constructing the index, it is possible to support queries based on a document structure as well as a logical inclusion between elements. In addition, the parsed image information is transported into the content-based indexer in order to get the index information of its color and its shape. To obtain an image feature vector for shape, it is possible to use the image object produced by the preprocessing part and generate a high-dimensional feature vector based on distances between the center point and a set of edge points. To generate a color feature vector, it is necessary to generate a color histogram and normalize the color histogram by dividing it by the number of pixels.

The structure-based and content-based index information is separately stored into their index structures, respectively. The index structures for structure-based retrieval are constructed by indexing XML documents based on an element unit and consist of keyword, structure, element, and attribute index structures. The index structure for content-based retrieval is a high-dimensional index structure, like X-tree (Berchtold, Keim & Kriegel, 1996) or CBF (Han & Chang, 2000), so as to store and retrieve both color and shape feature vectors efficiently.

Using the stored index information extracted from a set of XML documents, some documents are retrieved by the retrieval part in order to obtain a unified result to answer user queries. There is little research on retrieval models for integrating structure- and content-based information retrieval. To answer a document structure query, a similarity between an element q and an element t is computed as the similarity between the term vector of node q and that of node t by using a cosine measure (Salton & McGrill, 1983). Also, a similarity between an element q and a document D is computed as $\text{Max}\{\text{COSINE}(\text{NODE}_q, \text{NODE}_{Di}), 0 <= i < n\}$. To answer an image content query, it is possible to compute a similarity between the query feature vector and the image feature vector as 1 – (Euclidean distance / maximum distance) and retrieve relevant documents with high similarity in decreasing order of the similarity. In the case of content-based retrieval based on both color and shape feature vectors, a similarity measure is required. A similarity between a query image q and a target image t in the database is calculated as (1-Distc(q,t)/Nc)*(1-Dists(q,t)/Ns) where Distc (Dists) means a color (shape) vector distance between a query image and a target image, and Nc (Ns) means the maximum color (shape) distances for normalization.

Finally, a final document set that is acquired by integrating preliminary results from both structure- and content-based retrieval is given to users through a convenient user interface, such as a Web browser. To design an efficient structure- and content-based query interface, it is necessary to classify XML queries into two types: simple and composite. The simple query can be divided into keyword, structure, attribute, and image query. The composite query is the composition of simple queries, like structure plus keyword, structure plus attribute, image plus keyword, and image plus structure. It is shown from some experiments that it takes much more time to answer the structure query, compared to the other types of simple queries (Chang, 2002).

FUTURE TRENDS

Directions for future work can be studies on new information retrieval models for integrating preliminary results acquired from both structure- and content-based retrieval, because the two types of retrieval are very different each other, in terms of their nature and property. This can be achieved ultimately by trying to handle MPEG-7 compliant XML documents (Haoran, Rajan & Liang-Tien, 2003; Westermann & Klas, 2003).

CONCLUSION

For effective Web document retrieval, it is very important to retrieve XML documents based on both document structure and image content. To support structure-based retrieval, it is necessary to index XML documents based on the basic element unit, thus generating four index structures: keyword, structure, element, and attribute. To support image content-based retrieval, it is necessary to construct a high-dimensional index structure like CBF (Han & Chang, 2000) for retrieving both color and shape feature vectors efficiently.