IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB8036

Chapter VIII

A Multimedia Document Retrieval System Supporting Structure- and Content-Based Retrieval

Jae-Woo Chang and Du-Seok Jin
Chonbuk National University, South Korea

Recently it is common for users to acquire through the World Wide Web a variety of multimedia documents. As the number of Web documents is dramatically increasing, we need to develop a multimedia document retrieval system that can support both structure-based retrieval and content-based retrieval. In order to support structure-based retrieval, we design efficient index structures (i.e., keyword, structure, element and attribute) and implement those by using the o2store storage system. For the content-based retrieval, we implement high-dimensional index structure for color and shape feature that is based on X-tree. Finally, we do the performance evaluation of our multimedia document retrieval system in terms of system efficiency, such as retrieval time, insertion time and storage overhead, as well as system effectiveness, such as recall and precision measures.

INTRODUCTION

Because the number of Web documents is dramatically increasing, it is difficult for users to reach specific Web documents. Thus, it is necessary to develop a multimedia information retrieval system that can support both structure- and content-based retrieval. In 1996, XML (eXtensible Markup Language) was proposed as a standard markup language (W3C, 2000). An XML has flexibility of expression like SGML and is also easy to use like HTML.

In general, the conventional information retrieval systems for Web documents support only structure-based retrieval; it is difficult to deal with a user query efficiently. In this chapter, we design and implement a multimedia information retrieval system that can efficiently retrieve Web documents based on both document structure and image content. In order to support the structure-based retrieval, we design four index structures (i.e.,

keyword, structure, element and attribute) and implement those by using the o2store storage system. For the content-based retrieval, we implement high-dimensional index structure for color and shape features that are based on X-tree.

This chapter is organized as follows. In the first section, we introduce related works in the area of structure-based and content-based information retrieval. Folowing that we design a structure- and content-based multimedia document retrieval system. In the next section we present the implementation and performance evaluation of our system. Finally, we draw conclusions and provide some issues for future research.

RELATED WORK

Structure-Based Retrieval

Because an element is a basic unit that constitutes a structured document (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. Since there are a lot of studies on SGML documents, we, in this section, describe some related work on the representation of SGML document structures. First, RMIT in Australia proposed five query types for structure-based retrieval that should be supported in SGML information retrieval (Sacks-Davis, Arnold-Moore & Zobel, 1994). Most of the types consist of retrieval on upper-level elements (e.g., parent element), or on lower-level elements (e.g., child elements) from a given element. For supporting the five types of queries, RMIT proposed a *subtree model* which indexes all the elements in an SGML document and stores all the terms which are appeared in the elements (Lowe, Zobel & Sacks-Davis, 1995). Although the model supports efficient retrieval on a specific query, it has disadvantages of long indexing time and high storage overhead because index information should be repeatedly stored according to a tree depth. Secondly, RMIT proposed a *SCL structure* that extends the *GCL structure* (Dao & Sacks-Davis, 1996). After assigning numbers to both terms and markups in SGML documents, they use the *SCL structure* to store term interval, markups and inclusion relationships among elements. The *SCL structure* has an advantage that it can handle graph-structured documents, but it has two disadvantages that it requires a deletion operation and it cannot represent the depth of the elements effectively. Finally, SERI in South Korea proposed a *K-ary Complete Tree Structure* which represents a document as a K-ary complete tree. In this method, each element corresponds to a node in a K-ary tree. Therefore, a relationship between two elements can be acquired by calculation. This method has an advantage that it is fast to find an element including a given logical relation by calculation. But, as the depth of a K-ary tree is deeper, the number of nodes is increasing exponentially with a large number of unused nodes. In the cases of partial insert and deletion, almost all of nodes should be changed in their assigned number.

Content-Based Retrieval

There have been many researches on content-based retrieval techniques and a storage structure of multimedia DBMS. The key issues of the studies include keyword extraction for text-based retrieval, image-processing techniques used for feature extraction of images, and multidimensional indexing techniques for fast retrieval and content-based image retrieval based on color histogram, texture and shape. First, the *Query By Image Content*