XML문서를 위한 구조 및 내용기반 문서검색 시스템의 설계 및 구현

이정재, 장재우
전북대학교 컴퓨터공학과
jilee@dblab.chonbuk.ac.kr

Design and implementation of a structure- and content-based document retrieval system for XML documents

Jeong-Jae Lee, Jae-Woo chang
Department of Computer Engineering, Chonbuk National University

최근 XML문서에 대한 활용이 늘어나면서 이들 문서에 대한 저장 및 검색에 대한 요구가 증가하고 있다. XML문서는 SGML(Standard Generalized Markup Language)문서가 가지고 있는 다양한 기능들과 구조적인 표현 능력, 그리고 사용의 용이성 등의 장점을 지닌 언어로 1996년 웹의 문서 표준으로 제안되었다. 따라서 XML문서의 특성을 반영한 문서 검색시스템에 대한 요구가 시급한 상태이며, 기존의 시스템의 경우 구조 및 내용-기반 멀티미디어 문서검색을 효과적으로 지원하지 못하고 있다. 본 논문에서는 XML문서의 구조정보 및 내용정보를 효과적으로 검색할 수 있는 XML 문서 저장 시스템을 설계 및 구현한다. 구현하는 시스템은 구조-기반 검색을 위해 o2store위에 역파일 인덱스를 구축하고 내용-기반 검색을 위해 X-tree를 사용한다. 또한 검색 인터페이스를 JAVA로 구현하여 효율적인 검색이 이루어지도록 한다.

1. 서론

21세기 고도 정보화 사회에서 인터넷의 발달로 멀리 떨어진 사이트간에 텍스트뿐만 아니라 이미지, 오디오, 비디오를 포함하고 있는 멀티미디어 문서를 전달하는 것이 보편화되고 있다. 아울러 인터넷을 통해 전달되는 멀티미디어 문서의 개수가 기하급수적으로 증가함에 따라 사용자가 요구하는 멀티미디어 문서를 보다 효과적으로 저장 및 검색하는 것이 필수적이다. 인터넷을 통해 전달되는 멀티미디어 문서로는 HTML이나 XML 혹은 SGML과 같은 구조화된 문서 등이 있으며, 다양한 미디어를 포함하고 있다는 특징을 지닌다. XML은 SGML이 가지고 있는 다양한 기능들과 구조적인 표현 능력, 그리고 사용의 용이성 등의 장점을 지닌 언어로 1996년 웹의 문서 표준으로 제안되었다[1]. 따라서 XML문서의 특성을 반영한 문서의 저장 및 검색시스템에 대한 요구가 시급한 상태이며, 기존의 시스템의 경우 구조 및 내용-기반 멀티미디어 문서검색을 효과적으로 지원하지 못하고 있다[2].

따라서 본 연구에서는 인터넷에서 XML문서의 구조정보와 내용정보를 효과적으로 검색할 수 있는 XML 문서 검색 시스템을 설계 및 구현한다. 구현하는 시스템은 구조-기반 검색을 위해 o2store위에 구조 인덱스를 구축하고 내용-기반 검색을 위해 고차원 색인 구조인 X-tree를 사용한다[3].

2. 관련연구

2.1. 구조-기반 검색

SGML의 구조 인덱스 접근 방식에는 3가지가 있다[4]. 첫째는 K-ary Complete 트리 구조로서 이 방법은 엘리먼트 단위의 설계 방법이다. 둘째는 문서단위의 문서구문 트리 구조로서 문서단위의 설계방법이다. 셋째는 엘리먼트 단위의 문서구문 트리 구조로서 SGML parser를 통해 나온 구문 트리 정보를 엘리먼트 단위의 레코드들로 분리시켜 저장하는 방법이다. 세 가지 방법중 엘리먼트 단위의 문서구문 트리 구조 방법은 하나의 문서 구조 정보를 여러 개의 엘리먼트 단위 레코드로 저장하고 이를 빠르게 접근하기 위해 B+tree를 사용한다. 또한 부분 삽입과 삭제 시 우수한 성능을 보이고 부가 저장공간이 적은 장점을 지닌다.

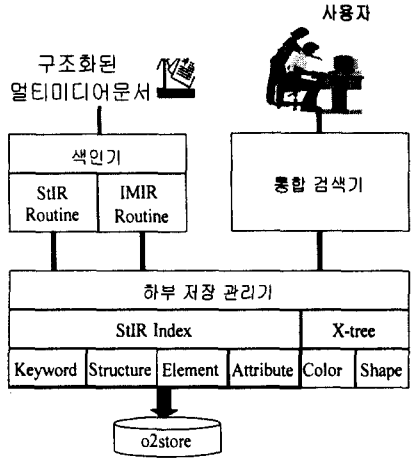
2.2. 내용-기반 검색

대용량 멀티미디어 데이터의 내용-기반 검색연구는 첫째 방대한 양의 이미지와 비디오 데이터베이스를 기반으로 텍스트 키워드, 색상, 모양, 질감, 구도 같은 다양한 검색질의를 제공하는 QBIC(Query By Image Content)시스템이다[5]. QBIC시스템은 미국 IBM Almaden연구소에서 개발한 시스템이며 이미지에서 추출한 특징 벡터를 R*-트리와 Filtering 기법을 사용하여 빠른 검색 결과를 보장한다. 둘째는 이미지내의 객체간의 공간 위치와 색상에 대한 질의를 처리할 수 있는 VisualSEEK시스템으로 미국 콜롬비아 대학에서 이미지에 대한 검색과 브라우징을 위한 툴(tool)로서 개발하였다[6]. 이 시스템의 목적은 이미지 검색에 효과적인 CBVG(Content-Based Visual Query)시

스템을 구현하는데 있다.

3. 구조 및 내용기반 문서검색시스템

XML문서는 SGML문서와 마찬가지로 추상화된 정보표현의 기본 단위가 엘리먼트이므로 기존의 정보검색에서의 문서 단위와 달리 엘리먼트 단위의 검색이 이루어져야 한다. 이러한 엘리먼트는 문서를 구성하는 논리적 구조에 의해서 정의되므로 문서의 내용에 의한 검색 이외에 문서 구조에 의한 검색이 필요하다. 인터넷에서의 XML문서 자체의 구조정보 및 내용에 기반한 특성을 고려한 문서 검색 시스템 아키텍처는 [그림 1]과 같다.



[그림 1] XML문서검색시스템 아키텍처

인터넷에서 접근할 수 있는 문서인 XML문서를 분석한 후 문서내의 구조정보와 내용에 기반한 이미지 정보를 추출하여 추출된 정보를 색인하는 색인기 부분, 데이터를 저장하고 접근하기 위한 하부 저장 관리기 부분, 저장된 데이터를 구조 및 내용-기반 검색하기 위한 통합검색기 부분으로 구분할 수 있다.

3.1. 색인기

먼저 색인기를 통한 문서 저장 과정을 살펴보면 엘리먼트 사이의 관계정보를 추출하기 위해서 XML문서는 고유의 DTD에 적합한 트리로 구성되어야 한다. 트리를 구성하기 위해서 Sp-1.3 parser를 통해 나온 파싱 데이터를 이용한다. 파싱 데이터를 이용하여 추출된 정보에서 전체 트리는 하나의 XML문서를 표현하고 각 노드는 하나의 엘리먼트를 나타낸다. 이러한 노드들의 정보들은 StIR(Structured Information Retrieval)루틴과 IMIR(Image Information Retrieval)루틴을 통해서 처리된다. 먼저 StIR 루틴을 보면 구조정보와 색인어 추출기를 통한 키워드들을 추출한다. 그리고 IMIR루틴은 이미지가 가지는 엘리먼트의 속성값과 객체식별자를 가지고 처리한다. 그 처리과정은 jpg포맷 이미지의 색상과 형태에 대한 특징벡터를 추출하기 위해서 비트맵 포맷 이미지로 변환한 후에 22차원의 색상 특징벡터와 24차원의 형태 특징벡터를 추출한다[7].

3.2 하부 저장 관리기

구조 및 내용-기반 멀티미디어 정보 검색 시스템을 위한 하부 저장 구조는 크게 두 가지 유형으로 설계되어진다. 첫째는 구조-기반 검색을 위해 역화일 기법을 이용하여 o2store 하부저장 시스템에 저장하는 것으로 StIR 루틴을 통해 전달된 정보를 Keyword, Structure, Element 그리고 Attribute의 인덱스로 구축한다. 둘째는 내용-기반 검색을 위해 고차원 특징벡터를 효율적으로 검색하고 k-최근접질의, 범위질의 등과 같은 다양한 검색 질의를 제공하는 X-tree를 이용하여 MIR루틴을 통해 전달된 Color와 Shape에 대해 특징벡터를 인덱스를 통해 저장한다.

3.3. 통합 검색기

검색과정에서 우선 검색기는 전달된 사용자 질의를 인덱스 관리자가 처리할 수 있는 형태로 재구성한다. 따라서 구조검색질의와 이미지 검색질의를 사용자가 함께 줄 경우 그 질의는 각 인덱스 관리자가 처리할 수 있는 값과 특징벡터로 재구성되어 사용자질의와 검색된 결과의 유사성을 계산하는데 먼저 내용질의 유사성은 다음과 같은 공식을 이용한다.

$$C_w = \begin{cases} \frac{Distc(q,t)}{Nc}, & \text{if a shape query is empty} \\ \frac{Dists(q,t)}{Ns}, & \text{if a color query is empty} \\ \frac{Distc(q,t)}{Nc} * \frac{Dists(q,t)}{Ns} \end{cases}$$

위 식에서 질의 이미지 q와 데이타베이스 내의 대상 이미지 t 사이의 유사도 CW(q, t)는 사용자 질의 이미지에 대한 하나의 검색된 이미지 사이의 가중치(Weight) 값을 나타내며, 이 값은 실제로 검색된 이미지와 함께 브라우징 된다. 여기에서 Distc(q, t)와, Dists(q, t)는 각각 질의 이미지 q와 대상 이미지 사이의 색상과 형태에 대한 거리를 계산한 값이다. Nc와 Ns는 각각 색상과 형태의 유사성을 정규화 하기 위한 값이다. 또한, 구조 질의의 엘리먼트 term 벡터간의 유사성은 다음과 같은 공식을 이용하여 계산한다.

$$S_w = COSINE(NODE_q, NODE) = \frac{\sum_{k=1}^{m}(TERM_{qk} \cdot TERM_{tk})}{\sqrt{\sum_{k=1}^{m}(TERM_{qk})^2 \cdot \sum_{k=1}^{m}(TERM_{tk})^2}}$$

구조 질의에서 노드q의 구조 질의 유사성은 노드t의 term벡터와 노드q의 term벡터 간의 유사성(Sw)을 나타낸다. 이러한 계산과정을 거쳐 출력된 검색 결과는 멀티미디어 문서 브라우징

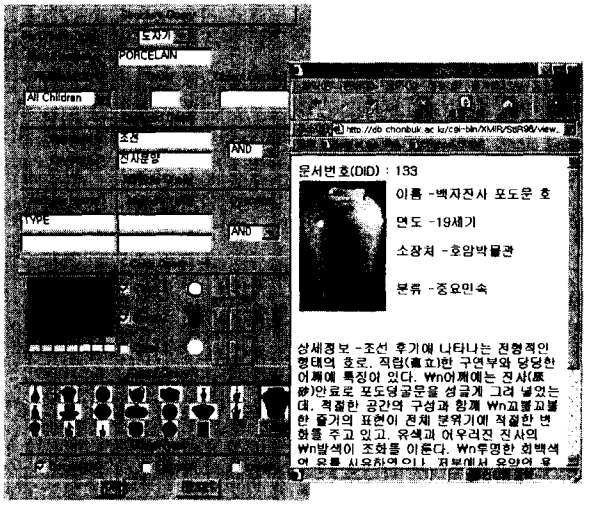
모듈을 통해 인터넷에서 브라우징 할 수 있는 형태로 합성되어 진다.

4. 응용분야 및 사용자 인터페이스

본 논문에서 사용한 XML문서로는 국립중앙박물관의 청자, 백자, 분청사기 도감에서 약 170여건의 문서를 XML DTD형식에 맞게 구현하여 사용하였다. 사용자는 원하는 XML문서를 검색하기 위해 구조적인 정보와 내용 정보에 관하여 검색 질의를 수행할 수 있다. 사용자 인터페이스에서 처리할 수 있는 질의의 유형을 분류하면 단순질의, 복합질의로 나눌 수 있는데, 단순질의에는 키워드, 구조, 엘리먼트, 애트리뷰트, 이미지칼라, 이미지색상를 줄 수 있다. 각 질의의 대표적인 예를 들면 아래와 같다.

- 키워드질의 : "귀현사기"가 있는 문서를 찾아라.
- 구조질의 : [도자기]엘리먼트의 자식 엘리먼트를 찾아라.
- 엘리먼트질의 : [museum]엘리먼트가 있는 문서를 찾아라.
- 애트리뷰트질의 : type이 백자인 문서를 찾아라.
- 이미지질의 : 특정 색상 또는 특정 형태를 가진 이미지를 찾아라.

복합질의는 "키워드 질의 + 이미지질의"와 같은 단순질의의 복합 형태들이다. 이미지 정보의 질의에 대한 검색에 있어서 원래 이미지를 축소한 아이콘 이미지를 출력하고 이를 바탕으로 사용자는 자신이 검색하고자 하는 아이콘 이미지를 선택함으로써 원하는 이미지에 대한 전체 정보를 검색할 수 있다.



[그림 2] 사용자 인터페이스

JAVA로 구현한 시스템의 사용자 인터페이스는 [그림 2]와 같다. 이는 문서의 시작엘리먼트를 porcelain으로서 그의 자식엘리먼트 전부를 찾는 경우이며, 키워드질의는 조선과 진사문양을 주었을 경우이다. 아울러 칼라는 흰색과 갈색의 비중을 8 : 2로 주고, 모양은 항아리모양을 선택한 복합질의의 인터페이스와 그 결과 문서를 보여준 예이다.

5. 결론 및 향후연구

본 연구에서는 인터넷에서의 다양한 멀티미디어 문서에 대한 구조 및 내용-기반 검색을 위해 XML문서를 이용한 문서검색 시스템을 설계 및 구현하였다. XML문서의 키워드 및 구조 정보의 저장을 위해서 엘리먼트 단위의 구문 트리 저장 구조를 사용하였고, 내용에 기반한 이미지 정보의 저장을 위해서는 고차원 색인 기법인 X-tree를 사용하였다. 그리고 구성된 인덱스의 저장 및 관리를 위해 o2store를 하부저장시스템으로 사용하였다. 따라서 XML문서의 복잡한 구조정보와 내용에 기반한 이미지 정보를 효율적으로 저장하고 검색할 수 있는 시스템을 설계 및 구현하였다.

향후 연구과제로는 하부저장시스템으로 사용된 o2store를 Public 저장 시스템인 Shore로 대체하고, 다양한 DTD에 대해 질의 처리가 가능한 시스템을 구현하는 것이다.

Reference

[1] Extensible Markup Language(XML), "http ://www.w3.org/TR/PR-xml-971208"

[2] R. Sack-Davis, T. Arnold-Moore and J. Zobel, "Database Systems for Structured Documents," Informational Symposium on Advanced Database Technologies and Their Integration, 1994

[3] S. Berchtold. D. A. Keim, H-P. Kriegel, The X-tree : An Index Structure for High-Dimensional Data, Proceeding of the 22nd VLDB Conference

[4] 손정한, 한성근, 장재우, 주종철, "SGML 정보 검색 인덱스 설계를 위한 K-ary트리, 문서 단위 구문 트리와 엘리먼트 단위 구문 트리의 비교", 98'한국정보과학회 가을 학술 논문집 Vol. 25. No. 2, pp.383-385, 1998

[5] W. Niblack, et. al., "The QBIC project: Quering by Image Content Using Color, Texture, and Shape," Proc. of SPIE Storage and Retrieval for Image and Video Databases, pp. 173-187, 1993

[6] J. R. Smith, S. F. Chang, "VisualSEEk: a Fully Automated Content-Based Image Query System," ACM Multimedia Systems, Nov 1996.

[7] Choon-Bo Sim, Kwang-Taek Song, Jae-Woo Chang, Joon-Whoan Lee, and Jae-Dong Yang, "Design and Implementation of a Content-Based Multimedia IR System for Cyber Museums", SPIE Electronic Imaging and Multimedia Systems II, pp 86-93 (1998).