구조화된 문헌의 이미지정보 검색시스템에 관한 연구
Design of an Image Retrieval System for Structured Documents

박현주, 정영미
연세대학교 문헌정보학과

Park Hyun-Joo, Chung Young-Mee
Dept. of Library and Information Science, Yonsei Univ.

전자도서관 환경에서의 필요성으로 인해 이미지정보에 대한 관심과 정보로서의 가치 및 중요성이 널리 인정됨에 따라, 이를 위한 효과적인 색인방법과 검색시스템이 요구되고 있다. 이에 본 연구에서는 구조화된 문헌의 이미지정보를 자동으로 색인하여 데이터베이스를 구축하고 WWW상에서 이용할 수 있는 이미지정보 검색시스템을 구현하였다.

1 서론

상당수의 전자문헌을 시스템 환경에 제한됨이 없이 처리하고 문헌의 논리적 구조를 나타낼 수 있는 국제적인 표준형식의 필요성이 대두되면서 구조화된 전자문헌을 기술하기 위한 여러 가지 국제 표준이 제정되었다. 한편 디지털 이미지 처리기술이 도입되면서 전자문헌에는 문자정보뿐만 아니라 이미지정보까지 포함할 수 있게 되었고, 특히 전자도서관 환경에서 조형물이나 건축물과 같이 물리적인 형태를 갖는 객체까지 수용하려는 노력으로 이미지정보의 가치와 중요성이 널리 인정되고 있는 실정이다.

이에 본 연구는 구조화된 문헌의 이미지정보를 자동으로 색인하여 데이터베이스를 구축하고 WWW상에서 검색시스템을 구현함으로써 막대한 양의 이미지정보를 효과적으로 색인하고, 전자도서관이 지향하는 '소유'에서 '접근'으로의 변화하는 패러다임과 맥을 같이 하여 이용자로 하여금 쉽게 검색할 수 있도록 하는 데 그 목적을 두고 있다.

2 이론적 배경

2.1 이미지정보의 개념과 속성

이미지란 어떤 객체의 외적 형태에 대한 표현으로서, 이와 관련된 용어에는 화상, 도형, 지도, 그림, 사진, 설계도, 표, 그래프, 삽화 등 여러 가지가 있다.

이렇게 이미지정보에 포함되는 개념과 종류는 다양하지만, 일반적으로 이미지정보는 본문의 내용을 효율적으로 보충해 주며, 객체 간의 관계나 비교 혹은 계층구조 등, 텍스트로는 여러 문단에 걸쳐 장황하게 설명해야 할 정보를 간단하고 명확하게 표현할 수 있는 매체이다.

이러한 이미지정보가 갖는 속성으로는 전기적 속성, 주제적 속성, 예시적 속성, 관계적 속성 등이 있다.

2.2 이미지정보의 색인 및 검색 특성

이미지정보에 대해 지시적·선별적 기능을 제공하기 위해서는 상세한 색인이 필요한데, 이미지정보의 색인과 검색에는 학제적 성격, 주관성 개입, 일관성 결여 등과 같이 이미지정보가 갖는 특성으로 인한 복잡한 문제가 필연

적으로 수반된다.

이미지정보의 색인 및 검색에 관한 연구는 처리대상에 따라 텍스트처리에 기반을 둔 연구와 화상처리에 기반을 둔 연구로 구분해 볼 수 있다. 기존의 화상처리에 기반을 둔 검색시스템에서는 이미지정보의 의미나 주제를 표현할 수 없는 문제가 있었으며, 텍스트처리에 기초한 연구는 제한된 어휘나 코드를 부여함으로써 적합한 이미지정보를 제대로 검색할 수 없는 한계가 있었다.

전자도서관 환경에서 물리적 형태를 갖는 객체도 수용하기 위해서는 객체에 대한 설명과 함께 그 특징을 적절하게 묘사하고 필요한 데이터를 제공할 수 있는 이미지를 포함시켜야 한다. 그러한 이미지는 다른 이미지나 텍스트, 혹은 다른 객체와 서로 연관될 수 있으므로, 동일한 이미지라 할지라도 주변상황이나 그 다음에 따라오는 이미지에 따라 그 의미가 변화될 수 있다. 따라서 이미지정보는 그 이미지가 속해있는 배경이나 문맥 내에서 해석되어야 하고, 색인과정에서 이러한 관계와 특성을 반영해 줄 필요가 있다. 또한 이미지정보를 색인할 때에는 이용자의 요구와 도서관의 효과적인 자료관리상의 필요성을 고려해볼 때, 문헌자료와 함께 처리할 수 있는 색인방법을 사용하는 것이 중요하다. 또한 이미지의 속성에 기반을 둔 접근점을 제공하고, 개별 이미지가 아닌 이미지의 집합에 대한 접근점을 제공해야 한다.

2.3 표준형식에 의한 문헌의 구조화

1) SGML문헌의 구조와 이미지정보의 표현

SGML 문헌은 SGML 선언부, 문헌유형정의부, SGML 문헌으로 구성된다.

이러한 SGML에서 이미지정보를 표현하는 방법에는 (1) 이미지정보를 직접 표현할 수 있도록 DTD를 설계하는 방법, (2) 표기법선언을 통해 특수하게 처리된 데이터를 포함시키는 방법, (3) 별도의 화일로 저장하여 참조하는 방법 등이 있다.

2) HTML문헌의 구조와 이미지정보의 표현

HTML을 따르는 문헌의 구문은 SGML의 응용으로서, HTML 3.2판에 따라 작성된 문헌의 경우 HTML의 다른 판으로 작성된 문헌과 구분하기 위해 반드시 <!DOCTYPE>선언으로 시작해야 하고, <HEAD>문헌요소와 <BODY> 문헌요소로 구성되는 <HTML>문헌요소가 그 뒤에 기술된다.

현재 HTML 3.2판에서 지원하는 이미지정보는 크게 <TABLE>문헌요소와 <IMG>문헌요소로 나누어 볼 수 있다.

2.4 WWW에서의 SGML 문헌의 활용

WWW서비스가 보편화됨에 따라 상당수의 기관이 인터넷 상에서 전자출판 등을 통해 정보를 배포하는 데 관심을 기울이게 되면서, 문헌의 구조를 충분히 표현할 수 있는 SGML문헌을 WWW에서 활용하는 문제에 관심을 기울이고 있다.

WWW에서 SGML문헌을 활용하는 방법은 (1) SGML문헌을 HTML형식으로 변환시켜 활용하는 방법, (2) SGML문헌을 직접 WWW에서 이용하는 방법, (3) SGML을 이용한 새로운 WWW용 마크업언어를 개발하는 방법 등과 같이 크게 세 가지로 구분해 볼 수 있다.

3 이미지정보 검색시스템의 구현

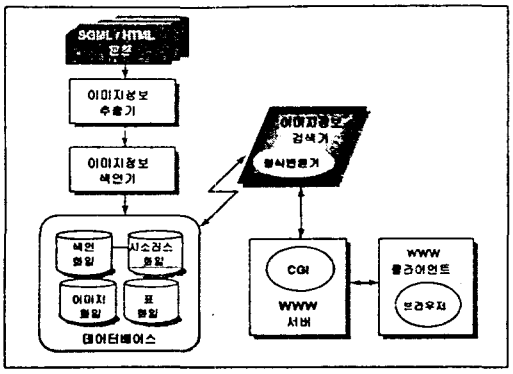
3.1 시스템 구현환경

본 시스템은 NCSA의 1.4.3버전의 WWW서버상에서 구현하였고, 브라우저로는 Netscape Navigator 3.0 Gold판을 이용하였다. 문헌의 형태소분석 및 색인을 위해 HAM 2.03판을 사용하였으며, 데이터베이스는 miniSQL 1.0.6판으로 구축하였다. 검색을 위한 CGI프로그램과 이미지정보의 추출 및 문헌형식 변환을 위한 프로그램은 Perl 5.001로 작성하였다.

실험대상 문헌으로는 「한국정보관리학회 논문대회 논문집」의 제1회에서 제3회까지 발표된 논문 가운데 16편을 선정하였다.

3.2 시스템 구성

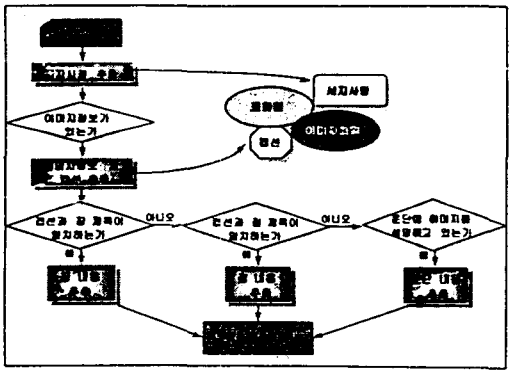
본 시스템은 크게 WWW서버와 WWW클라이언트, 이미지정보 추출기, 이미지정보 색인기, 데이터베이스, 이미지정보 검색기로 구성되며, 전체적인 시스템 구성은 <그림 1>과 같다.



<그림 1> 이미지정보 검색시스템 구성

1)이미지정보 추출기

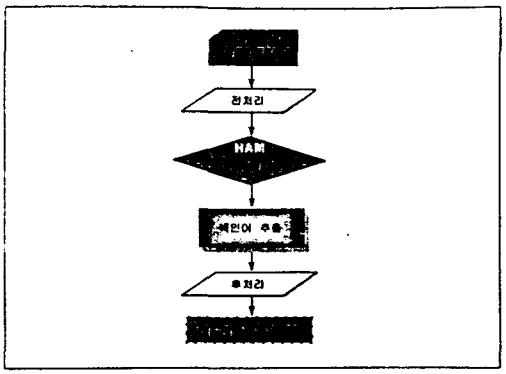
본 이미지정보 추출기는 입력된 문헌에서 출현하는 이미지정보와 이미지정보를 설명하고 있는 텍스트, 캡션, 서지사항을 각각 별도의 화일로 추출해낸다. 이러한 이미지정보 추출과정을 전체적으로 나타내보면 <그림 2>와 같다.



<그림 2> 이미지정보 추출과정

2) 이미지정보 색인기

이미지정보 색인기에서는 먼저 전처리 과정을 통해 이미지정보 추출기에서 추출된 텍스트에서 태그를 모두 제거한 후, HAM을 이용하여 조사나 어미 및 불용어를 제외한 대부분의 명사를 색인어로 추출하고, 후처리 과정을 거쳐 데이터베이스에 입력된다. 이미지정보 색인기의 전체적인 구성은 <그림 3>과 같다.



<그림 3> 이미지정보의 색인어 추출과정

3) 데이터베이스

본 시스템의 데이터베이스는 miniSQL을 사용하여 구축한 색인화일 및 시소러스화일, 그리고 한글 vi에디터로 입력된 표화일과 GIF형식의 이미지화일로 구성된다.

색인화일과 시소러스화일의 레코드 구조는 각각 <표 1>, <표 2>와 같다.

<표 1> 색인화일의 레코드 구조

입력 번호	색인어	이미지 정보	캡션	서지 사항	논문 번호

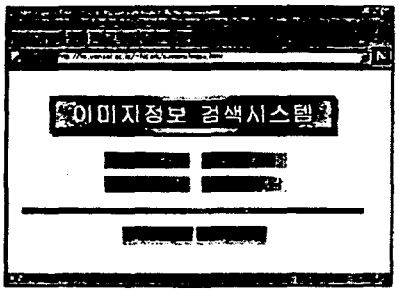
<표 2> 시소러스화일의 레코드 구조

색인 어	상위 개념어	하위 개념어	관련 어	우선 어	비우 선어

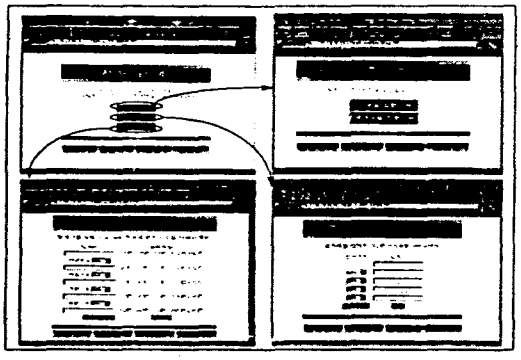
4) 이미지정보 검색기

본 연구에서 구현한 이미지정보 검색기는 WWW서버와 데이터베이스를 연결하는 CGI프로그램과 이미지정보 검색시스템의 WWW페이지로서, 이미지정보 검색시스템의 첫페이지는 <그림 4>와 같고, 이미지정보의 검색기능에는 <그림 5>와 같이 집합검색 기능, 단순검색 기능, 확장검색 기능 등 3가지 검색기능이 포함된다.

<그림 4>



검색시스템의 첫페이지



<그림 5> 이미지정보 검색기능 선택화면

이미지정보 검색기의 CGI프로그램에는 데이터베이스에 저장되어 있는 SGML형식의 표를 HTML형식으로 변환시켜주는 문헌형식변환기가 포함되어 있으며, 이 형식변환과정은 검색된 이미지정보가 표일 때에만 이루어지게 된다. 또한 초기검색결과에서 정보요구에 적합하다고 판단되는 이미지정보를 선택하면 최종검색결과로 적합한 이미지정보 및 그 캡션과 함께 서지사항이 출력된다.

4 시스템 평가 및 결론

본 연구에서는 이미지정보를 포함하고 있는 구조화된 문헌의 계층구조를 이용하여 이미지정보를 자동으로 색인하고, 구축한 데이터베이스와 연결하여, 다양한 검색기능을 제공하는 이미지정보 검색시스템을 WWW상에서 구현하였다.

구현된 시스템을 이용한 검색실험에서는 확장검색이 단순검색보다 더 나은 검색능력을 가져왔으며, 확장검색기능은 텍스트의 길이가 긴 경우 이미지정보가 효과적으로 검색되었다. 또한 색인에 사용된 텍스트의 유형은 단순검색기능의 검색결과에 영향을 미치지 않았으나, 절이나 장을 이용하여 색인된 이미지정보는 확장검색기능 이용시 더 정확하게 검색되었다.

본 이미지정보 검색시스템의 특징으로는 이미지정보의 색인과정에서 문헌의 구조를 활용하여, 하나의 언어로 이미지정보 및 그와 관련되는 문헌을 동시에 검색한 점을 들 수 있다. 또한 WWW에서의 활용시에 SGML형식의 문헌을 HTML형식으로 변환하였고, 이미지정보를 집단화시켜서 검색할 수 있는 기능을 제공하는 한편, 단순검색기능이외에도 탐색어의 확장을 통한 검색기능을 제공한다.

참고문헌

1. Baxter, Graeme & Anderson, Douglas. "Image Indexing and Retrieval: Some Problems and Proposed Solutions," *New Library World*, 96(1123), 1995, pp.4-13.
2. Cole, Timothy & Kazmer, Michelle M.. "SGML as a Component of the Digital Library," *Library Hi Tech*, 13(4), 1995. pp.75-90.
3. Furuta, R., "Documents in the Digital Library," *Proceedings of International Symposium on Digital Libraries 1995*. (Tsukuba, Japan: University of Library and Information Science, 1995), pp.105-111.
4. Layne, Sara Shatford, "Some Issues in the Indexing of Images," *JASIS*, 45(8), 1994. pp.583-588.
5. Schamber, Linda, "What is Document? Rethinking the Concept in Uneasy Times," *JASIS*, 47(9), 1996, pp.669-671.