PDF를 이용한 학위논문 데이터베이스 구축

The PDF Database for Theses and Dissertations

임 상 원 (방송대학교)
최 원 태 (건국대학교)

Yim, Sang Won (Korea National Open University)
Choi, Won Tae (Kon-Kuk University)

전문 데이터베이스를 구축하는 방법은 이미지, ASCII, HTML, SGML, PDF 등으로 구분할 수 있다. 특히 PDF는 구축이 용이하고 제한된 기능이기는 하지만 비교적 저렴한 비용으로 전문 데이터베이스를 구축할 수 있다. 본 연구에서는 우리말의 지원이 가능한 Zwon사의 PDF 포맷을 이용하여 학위논문을 대상으로 데이터베이스를 구축하였다.

1. 서론

PDF(Portable Document Format)는 어도비(Adobe)사에서 일반 및 인터넷에서 통용되는 공통표준문서로 제안한 문서 파일 형식이다. PDF는 인터넷 이전에 이미 컴퓨터의 기종, 운영체제, 프린터 종류 및 해상도 등에 제한없이 어디서나 읽고 출력할 수 있도록 고안된 문서 형태이다. 그러므로 PDF는 인터넷에서 종이 인쇄물과 같은 역할을 하면서 전자문서 전자출판을 서비스 할 수 있는 솔루션이라 할 수 있다.

현재 우리 나라에서는 어도비사의 Acrobat PDF의 경우 한글 처리가 원할하게 지원되지 못하기 때문에 널리 사용되지 못하고 있다. 본 연구에서는 우리말의 지원이 가능한 Zwon사의 PDF 포맷을 이용하여 학위논문을 대상으로 데이터베이스를 구축하였다.

2. PDF 개요

PDF는 포스트스크립트(Postscript) 언어에 기반을 두고 만들어졌으며 자체의 압축기능을 포함하고 있어 인터넷/인트라넷에서 작은 파일 사이즈의 문서로 만들어 전송한다. 또한 On-line 환경이나 Off-line 환경에서도 여러 전송 수단을 통하여 문서 정보의 공유 및 전송 등의 여러 장점을 가지고 있는 File Format이다. PDF 파일은 화면에서 곧바로 선명한 화질로 내용을 볼 수 있음은 물론 프린터로 인쇄했을 때에도 깨끗한 인쇄물을 얻을 수 있다

HTML은 텍스트, 그래픽 등의 파일이 따로 존재하고 파일에 링크를 통하여 화면에 보여진다. 그러나 PDF는 파일 내에 모든 텍스트, 그래픽 등이 정해진 위치에 존재하고 폰트나 그래픽 등은 벡터(vector) 기반에서 그대로 보여주기 때문에 Reader에서 확대를 하거나 축소를 해도 폰트나 그래픽은 그 해상도가 변하지 않고 원래의 그래픽을 그대로 유지할 수 있는 고품질 출력물을 얻을 수 있다. PDF는 이미지가 아닌 Vector Graphic(line, point, area)을 지원하고 처리함으로서 zoom-in과 zoom-out시 고선명의 화면을 제공할 수 있다.

PDF는 페이지 단위로 제작되며 화면에서도

페이지 단위로 보여준다. 또한 오디오(QuickTime), 동영상(AVI movie)등의 멀티미디어 환경도 지원 가능하며 폰트 타입(type 1, truetype)을 PDF 내에 포함하고 있다.

3. DocuCom 개요와 기능

3.1 DocuCom의 개요

DocuCom은 대만의 Zwon사에 의하여 개발된 제품으로 DBCS(double bytes character set)을 지원한다. DocuCom은 DocuDriver(다른 응용 프로그램의 출력을 PostScript 파일로 생성), DocuMaker(Postscript 파일을 PDF 문서로 변환), DocuPlus(PDF 문서의 검색, 주석 첨부, 문서 편집에 사용), DocuReader(브라우저로서 모든 PDF 문서의 검색에 사용하며 무상으로 복사, 배포 할 수 있음) 등으로 구성되어 있다.

PDF를 열람할 수 있는 방법은 2가지이다. 첫째, Adobe Acrobat Reader를 통하여 온라인(인터넷/인트라넷)상에서 웹 브라우저에 Plug-In하여 PDF를 열람하는 방법이다. 둘째, Acrobat Reader 자체 Viewer 프로그램을 통하여 PDF를 볼 수 있는 방법이다. PDF는 일반 HTML과 마찬가지로 전체 PDF 파일을 전송 비율만큼 자료를 보여 주기 때문에 사용자는 전송 중에도 현재의 PDF 전송율만큼 자료를 볼 수 있다.

3.2 PDF 주석 기능

PDF 문서의 주석 작성은 문서와 멀티미디어의 편집, 관리, 검색이 가능하며 하이퍼 링크, 북마크, 마커, 미디어 클립, 노트, 아티클, 썸네일 등의 기능을 지원하고 있다.

북마크는 찾기 힘든 또는 자주 보는 정보를 찾는데 사용하며 특정 페이지 검색, 아티클 보기, PDF파일 또는 실행파일 열기, 인터넷 연결 등의 액션을 가질 수 있어 정보검색에 편리하다.

하이퍼링크는 문서와 문서의 연결, 특정 단어 또는 문장을 참조나 설명과 연결할 때 주로 사용하며 북마크와 같은 액션을 가질 수 있다. 특히 아이콘도 링크로 사용할 수 있어 검색의 편리성을 향상시킨다. 링크로 주석 불러내기 기능을 사용하면 의견이나 설명을 기록한 노트 또는 멀티미디어 자료나 다른 문서를 가지고 있는 미디어클립을 불러내어 작동시킬 수 있다.

텍스트 마커는 중요한 단어나 문장을 형광펜으로 강조하듯이 문서의 내용을 강조하는 데 사용하며 북마크와 같이 액션을 부여하여 하이퍼 링크로 사용할 수도 있다. 또한 주석 불러내기 기능도 지원한다.

미디어클립은 동영상, 사운드, 이미지 또는 다른 PDF 문서를 현재의 문서에서 별개의 창으로 볼 수 있다. 하이퍼링크나 텍스트 마커로 작동할 수도 있다.

아티클은 신문, 잡지의 박스기사와 같이 본문의 내용과 다른 주제를 같은 면에 구성할 수 있어 다양한 문서 구성이 가능하다. 아티클은 아티클 이름 지정이나 북마크, 하이퍼링크, 마커 등의 액션으로 검색할 수 있다.

노트는 의견, 설명 등을 기록하여 본문의 내용을 훼손하지 않고 문서에 첨부하는 기능으로 부서간의 검토의견, 내용에 대한 보조설명 등으로 사용한다. 부서간의 문서 교환이나 리포트에 대한 평가의 목적으로 사용할 수 있다.

썸네일은 문서 페이지 단위의 축소판으로 검색이나 문서편집에 사용할 수 있다.

4. 학위논문 PDF 구축 사례

4.1 워드프로세서 파일의 PDF 변환

본 연구에서는 우리나라에서 많이 사용되는 한글과 컴퓨터사의 흔글, 마이크로소프트사의 MS Word, 삼성의 훈민정음으로 구성된 데이터를 사용하였다. MS Word와 훈민정음은

PDF 파일로의 변환이 원활하게 지원되므로 변환시 데이터의 손실은 없다.

그러나 한글에서는 출력시 비트맵 폰트를 사용하므로 폰트가 이미지 처리되며 해상도는 떨어진다. 이러한 문제를 해결하려면 한글을 마이크로소프트 워드로 변환하여 PDF로 변환하면 해결되나, 한글에서 작성된 문서내에 그리기 모양으로 작성된 선과 모양은 워드에서 읽지 못한다. 위의 문제는 마이크로소프트에서 개발한 ara96cnv.exe 변환 프로그램을 인스톨 후 읽으면 해결되나 한글에서만 지원하는 한글 서체는 변환하여도 폰트가 깨어진다. 한글에서 윈도우에서 지원하는 폰트로 변환 후 워드로 읽어들이면 된다.

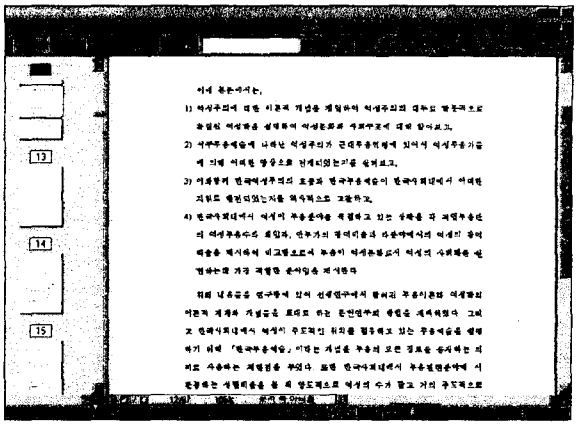

그림 2. PDF Sum 화면

4.2 PDF, SGML, HTML 구축 사례 비교

학위논문은 논문의 앞부분(표제면, 논문제출서, 논문인준서, 감사문, 헌사, 저작권페이지, 내용목차, 그림목차, 표목차, 부록목차, 국문초록, 키워드 등), 본문(장, 절, 항, 목 및 단락), 논문의 뒷부분(참고문헌, 부록, 색인, 용어집, 영문초록 등)으로 구성되어 있다(조왕근 1997).

본 연구에서 학위논문을 대상으로 구성된 HTML, SGML, PDF 구축 사례는 다음과 같다.

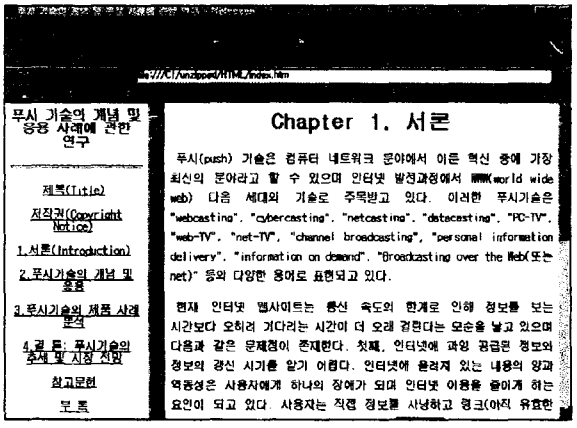

그림 3. HTML 화면

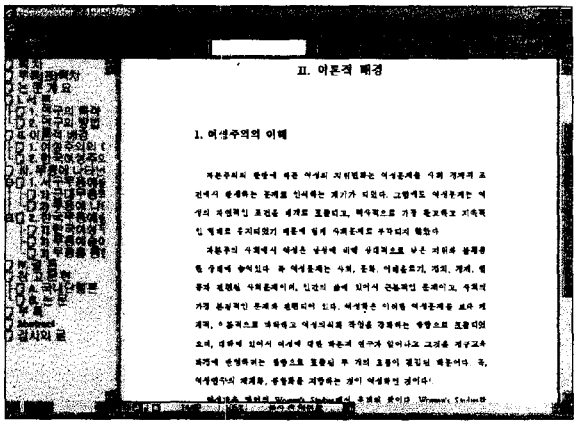

그림 1. PDF bookmark 화면

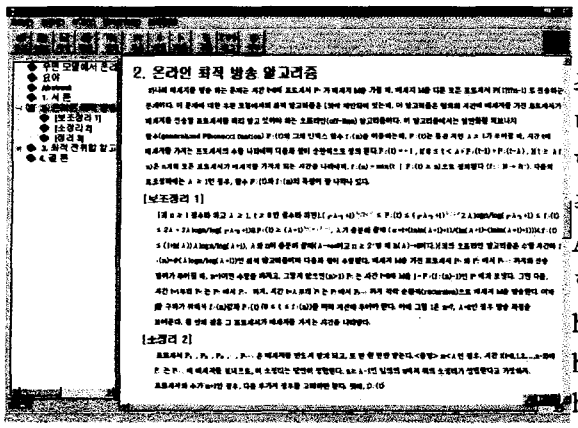

그림 4. SGML 화면

4. 결론

전문 데이터베이스를 구축하는 방법은 이미지, ASCII, HTML, SGML, PDF 등으로 구분할 수 있다. 이미지 방법은 구축 방법은 간단하나 전문의 검색을 지원하지 못하는 단점이 있다. ASCII 방법은 전문 검색은 가능하나 멀티미디어 처리에 제한이 있다. HTML 방법은 검색, 웹 브라우저의 보편화의 장점이 있으나 문헌의 구조 표현에 있어서 일반적인 단점이 있다. SGML은 문헌의 구조를 반영하므로 전문 데이터베이스의 구축에 표준화된 포맷으로 사용되어 왔다. 그러나 보편화된 DTD 에디터, 브라우저의 제한이 있으며 구축 비용이 고가인 단점이 있다.

PDF는 구축이 용이하고 제한된 기능이기는 하지만 비교적 저렴한 비용으로 손쉽게 주어진 기능을 수행할 수 있다. 그러나 국제적인 표준화, 검색의 제한성이 존재한다. 그러나 검색의 제한성은 텍스트를 처리하는 필터의 개발로 인하여 해결될 수 있다(최원태 1998). 이러한 필터를 이용하면 키워드를 입력하여 특정한 PDF 파일을 직접적으로 선택하고 키워드가 존재하는 해당 페이지의 직접적인 브라우징이 가능하다.

참고문헌

조왕근 1997. 학위논문 SGML DTD 및 데이터베이스 구축에 관한 연구. 한양대학교 교육대학원 석사학위논문.

최원태 1998. "A Digital Library Protype for Access to Diverse Collections," 한국문헌정보학회지, 32(2), 1998 : 295-307.

http://www.snikorea.co.kr

http://www.zeon.com

http://www.dki.co.kr