인터넷을 기반으로 한 DVI 포맷의 복합문서 전송 및 전문 데이터베이스 구축 사례 연구

윤화묵, 김진숙, 이기호
{hmyoon, ghlee}@ns.kordic.re.kr, jskim@bulls.kordic.re.kr
연구개발정보센터

Manipulation of Complex Documents of DVI Format in the Internet Environment and Construction of Full-Text Database

Wha-Muk Yoon, Jinsuk Kim, Kyi-Ho Lee
Korea Research and Development Information Center

요 약

1990년대 중반부터 인터넷의 활성화와 다양하고 강력한 문서편집기의 보편화에 따라 복잡한 문서들이 대량으로 생산됨에 따라 인터넷을 통한 효율적인 문서교환의 필요성이 늘어나고 있다. 그러나 생산된 방대한 양의 전자형태 복합문서들은 아래아한글, MS-Word, LaTex 등 다양한 문서편집기로 작성되었고 문서형식의 표준화가 이루어지지 않아, 효율적으로 활용되지 못하고 특히 문서교환에 있어 많은 문제점을 야기하고 있는 실정이다. 본 논문에서는 다양한 형태로 존재하는 복합문서들을 하나의 통일된 중간포맷으로 변환하고, 변환된 복합문서들을 전문데이터베이스(full-text database)화하여 이를 인터넷을 통해 효율적으로 검색할 수 있는 전문검색시스템 모델을 제시한다.

I 서론

인터넷의 사용이 날로 증가하면서 복잡한 문서 교환의 요구가 점차 증가하고 있다. 국내의 경우 표준 문서 작성기로 편집된 많은 문서들이 일부는 HTML과 같은 형태로 재편집되고 있으나 많은 경우 사장되고 있음으로 인한 정보의 손실이 매우 큰 실정이다. 더우기 학위논문과 같은 중요한 자료들이 전자형태의 문서가 이미 존재함에도 인터넷 전송이 어려운 이유로 인해 TIFF와 같은 이미지 형태로 인터넷을 통한 교환을 위해 다시 제작되고 있어서 경제적인 손실 또한 지대하다.

특히 과학기술분야의 학위논문과 같이 수학식이나 과학식이 포함된 문서의 경우 현재로서는 인터넷을 통한 일반적인 전송이 매우 어려운 형편이다. 수학식이나 과학식 처리의 요구가 높아지면서 이미 영어권에서는 JAVA애플릿이나 CGI프로그램, 플러그인과 같은 여러 방법 및 모델이 인터넷 전송의 방편들로 제시되고 있으며 PDF 포맷의 문서가 널리 통용되고 있다.

그러나 이러한 방법들은 인터넷 상의 전송이 효율적이지 못하고 한글을 비롯한 비영어권 문서의 처리에 한계가 있다. 이러한 문제점들을 해결하기 위해 본 연구에서는 문서를 인터넷 상에서 효과적으로 전송할 수 있는 문서 형식인 DVI 문서로 변환하는 모듈과 이미 변환된 DVI 문서를 정보 검색에 활용할 수 있게 하는 텍스트 추출 모듈 그리고 DVI 문서를 인터넷에서 효과적으로 전달 하기 위한 DVI 분할 모듈을 개발하였다.

2. 관련 동향

지금까지 복합문서와 관련하여 사용할 수 있는 도구 프로그램들은 단순히 문서를 HTML로 변환하는 것이 대부분이며 현재 영어권에서는 복합문서 전송의 한 방편으로 아도비(Adobe)사의 PDF가 널리 사용되고 있으며 사실상의 표준으로 받아들여지고 있는 형편이다.

이 외에도 아래아한글, MS 워드, 워드퍼펙트 등과 같은 워드프로세서들은 이들로 작성한 문서를 HTML로 변환하여 주는 도구 프로그램이나 CGI, 플러그인을 개발하여 제공하고 있다.

그러나 학위논문이나 연구논문과 같이 복잡한 수준의 문서는 HTML로 변환해서는 원래의 형식을 살리기가 사실상 불가능하며, PDF의 경우 문서파일의 내용을 이해하는 것이 힘들고 한글과 같은 비영어권의 2바이트 문자의 표현이 고려되지 않았다는 단점이 있다. PDF의 경우 한글을 처리하기 위해서는 한글 서체를 포스트스크립트 서체로 변환하여 내부에 담기 때문에 파일의 크기가 수 메가바이트에 달하게 되어 인터넷을 통한 한글 문서 교환용으로는 부적합하다.[5]

인터넷 상에서 복합문서를 전송하기 위해서는 원본과 동일하면서도 문서의 크기가 작아야 하고 또한 문서 내에서 원하는 부분만을 볼 수 있어야 한다. 이렇게 함으로써 사용자에게 원하는 정보를 빠르게 전달할 수 있을 뿐만 아니라 인터넷의 트래픽을 대폭 줄일 수 있다.

DVI형식은 학술 문서교환 및 출력을 위한 용도로 미국 스탠포드대학에서 고안된 문서 포맷이다. DVI는 학술 문서를 작성하기 위해 고안된 TeX의 기본 저장 형식으로 채택되면서 과학기술분야에서는 이미 오랫동안 문서 교환용으로 사용되어온 형식일 뿐만 아니라 문서가 원본과 동일하면서도 파일의 크기가 매우 작다는 장점을 지닌다. 또한 DVI에 관한 기술은 공개되어 있으므로 쉽게 정보시스템에 적용할 수 있다는 장점도 있다. 이러한 이유로 본 연구에서는 인터넷을 통한 복합문서 전송의 원형으로 DVI를 채택하였다.[1]

3. DVI 포맷 처리 도구

본 연구에서는 웹을 통한 DVI 복합문서 전송에 있어서 개발된 도구는 모두 네가지이다. 첫째 DVI 문서작성기는 윈도우즈 환경에서 각종 문서를 프린터 드라이버를 통해 DVI 문서로 변환한다. 둘째 DVI 분할기는 DVI 문서를 페이지 단위로 분할하여 사용자에게 전송한다. 셋째, DVI 본문 추출기는 페이지 단위로 본문의 텍스트를 추출한다. 마지막으로 DVI 뷰어는 사용자가 전송된 DVI 문서를 볼 수 있도록 한다.
그림 1.에서 이러한 네가지 기능을 이용하여 DVI 데이터베이스를 구축하는 과정을 볼 수 있다. 본 연구에서는 정보검색시스템으로 연구개발정보센터에서 개발한 KRISTAL-II를 이용하였다.[2]

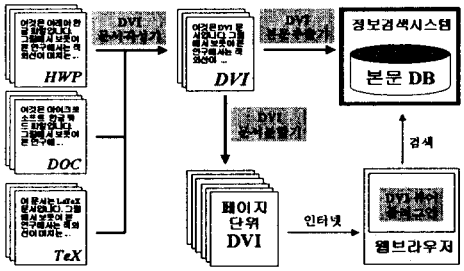

그림 1. DVI 문서 데이터베이스 구축

3.1 DVI 문서작성기

3.1.1 TeXplus 라이터

TeXplus 라이터는 마이크로소프트 윈도우 95/98/NT 상에서 한글과 컴퓨터사의 아래아한글을 제외한 응용 프로그램에서 DVI화일을 생성시켜주는 프린터 드라이브 소프트웨어다. 특히 우리나라에서 많이 사용하고 있는 한글 워드프로세서인 ㈜마이크로소프트의 한글워드, 삼성전자의 훈민정음, 핸디소프트의 아리랑 등에서 원본과 동일하게 수식과 그래픽을 포함한 한글 DVI파일을 만들 수 있도록 구현하였다. TeXplus 라이터는 TeX이 어려워서 사용하기 힘들었던 점을 고려하여 윈도우용 응용프로그램에서 쉽게 사용할 수 있도록 하였다. 응용 프로그램안의 인쇄 메뉴에서 프린터를 선택하듯이 TeXplus 라이터를 선택하여 출력하면 DVI형식으로 쉽게 변환할 수 있다.[3]

3.1.2. TeXplus HWP 라이터

TeXplus HWP 라이터는 TeXplus 라이터와 마찬가지로 마이크로소프트 윈도우 95/98/NT상에 설치된 HWP 파일에서 DVI화일을 생성시켜 주는 프린트 드라이브 방식의 소프트웨어다. HWP 문서는 한글워드나 엑셀, 훈민정음 등이 일반적으로 사용하는 윈도우 표준 인쇄 방식을 사용하지 않고 HWP 자체의 인쇄 경로를 사용하기 때문에 TeXplus 라이터를 사용하여 HWP 문서를 DVI로 변환할 수 없기 때문에 별도로 개발하게 되었다. TeXplus 라이터와 마찬가지로 TeXplus HWP 라이터 또한 원본과 동일하게 수식과 그래픽을 포함한 한글 DVI파일을 만들 뿐 아니라 텍스트를 코드값으로 저장하여 검색엔진을 쓰는 전자도서관 구축에 적당하다.

3.2 DVI 문서 분할

일반적으로 복합문서의 특징은 HTML 파일에 비해 파일의 크기가 매우 크다. 따라서 사용자들이 인터넷을 통하여 논문을 검색할 때 속도와 효율성 문제가 발생한다. 현재 웹상에서 복합 문서를 보기 위해서는 문서 전체를 모두 전송 받은 후 화면에 문서의 첫 페이지가 출력되기 때문에 문서가 클 경우 본문을 보려면 상당한 시간이 필요하다. 특히 논문 검색처럼 문서 전체를 읽으려는 목적이 아니라 문서의 일부만 보기를 원할 때에는 매우 비효율적이다.

이러한 속도 및 비효율성 문제를 개선하기 위해서는 첫째 원본 문서 자체의 크기가 작아야 하고, 둘째 원하는 부분만을 볼 수 있어야 한다. 본 연구에서는 이러한 필요성에 따라 DVI 문서를 페이지 별로 재구성하여 요구하는 부분만 전송하도록 했다. 즉 여러 페이지로 구성되어 있는 DVI 파일에서 필요로 하는 본문만을 떼어내서 새로운 DVI 파일로 재구성한 후 사용자에게 전송하는 방식이다. 이러한 작업을 해주는 도구가 바로 DVI 분할기(DVISPLIT)로 웹 환경에서 CGI로 구현되었다. 따라서 수백쪽에 이르는 DVI파일에서 특정 페이지를 삽입된 그림과 함께 추출하여 사용자의 검색 및 질의요구에 따라 특정 페이지만을 웹에서 제공할 수 있다.[4]

3.3. DVI 본문추출기

TeX의 중요한 실행 결과물인 DVI 파일의 내용을 보려면 xdvi 같은 DVI 뷰어기 필수적이며, DVI 파일을 텍스트로 변환해야 할 필요도 있다. 때문에 DVI 파일을 입력으로 해서 텍스트로 변환하는 프로그램은 이미 많이 나와 있다. 그러나 이들 프로그램들은 영어권에서 개발되었기 때문에 한글과 같은 2바이트 문자에 대한 고찰이 전혀 이뤄지지 않았다. 따라서 본 연구에서는 한글로 된 DVI 파일을 텍스트로 변환하는 프로그램을 작성하였다.

DVI 본문추출기는 TeXplus 라이터, TeXplus HWP 라이터 또는 TeX 문서가 만든 DVI 파일로부터 한글을 포함한 텍스트를 추출하여 주는 도구이다. 본문추출기는 하나의 문서에 대하여 페이지 단위로 각 페이지에

대한 정보와 수식이나 그림을 제외한 본문을 텍스트로 추출한다. 이렇게 추출된 텍스트는 정보검색시스템에 적재하여 문서 DB로 저장할 수 있다. 사용자들은 이 DB를 검색함으로써 DVI 문서를 페이지 단위로 검색할 수 있게 된다.

3.4 DVI뷰어(TeXplus 뷰어)

TeX출력 파일인 DVI문서를 인터넷에서 볼 수 있게 해 주는 플러그인으로서 별도의 프로그램을 실행할 필요가 없다. DVI파일을 열 때마다 실시간으로 실행되며, TeXplus 라이터로 생성한 DVI파일은 물론이고, 이외의 모든 TeX 컴파일러에서 생성한 DVI파일은 모두 TeXplus 뷰어를 통해 볼 수 있다. 윈도우즈 95/98/NT 환경에서 Netscape Navigator, Microsoft Internet Explorer 등과 같은 웹브라우저를 안정적으로 지원한다.

4. DVI 문서 검색서비스 구현 사례

국내 대학에서는 몇 년전부터 석박사 학위논문을 책자와 더불어 전자형태로 제출하도록하는 곳이 늘어나고 있다. 이러한 전자형태의 학위 논문은 아래아한글, MS WORD 및 TeX과 같은 3종의 문서편집기로 작성된 것이 대부분이나 현재까지 인터넷을 통하여 효율적으로 제공하고 있는 곳은 거의 없는 실정이다. 본 사례에서는 국내 2개 대학이 소장하고 있는 전자형태의 석박사 학위논문을 동일한 형식(DVI형식)으로 변환하여 DB를 구축하였다. 각 대학이 소장하고 있는 전자형태 학위논문의 분포는 표 1과 같다.

학위논문 서지사항은 각 대학이 보유하고 있는 서지정보를 제공받아 연구개발정보센터의 검색시스템인 KRISTAL-II에 적재하여 학위 논문을 검색하기 위한 자료로 활용하고 있다.[6]

본 연구에서는 각기 다른 형식으로 저장되어 있는 전자문서를 문서변환기(TeXplus 라이터, TeXplus HWP 라이터)를 이용하여 DVI 형태로 변환하였다. 이렇게 생성한 DVI파일은 학위논문 검색을 위하여 저장장치에 보관하게 된다. DVI파일은 서버의 적당한 위치에 저장하고, 색인 파일로 재구성하여 저장하게 된다. 페이지 단위로 추출된 텍스트 정보는 문서적재를 위하여 필요한 스키마를 작성함으로써 정보검색시스템 KRISTAL-II에 적재한다. 그림 2.는 DVI 형태의 학위논문을 검색하기 위한 전체적인 시스템 구성도를 보여주고 있다.

표 1. 전자형태 학위 논문의 문서편집기별 분포

문서형태	HWP	MS-Word	Tex	계
KAIST	617	863	431	1911
포항공대	110	136	79	325
계	727	999	510	2236

본 연구에서는 DVI 형태의 학위논문 2,200권과 이 논문들을 대상으로 텍스트전문 DB를 구축하였다. 사용자는 학위논문 서지정보 DB를 검색함으로써 각 학위논문의 초록이나 DVI형태의 원문 전부를 볼 수 있으며,

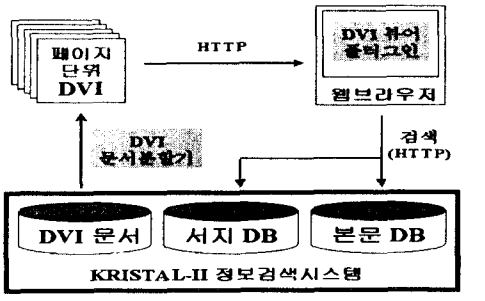

그림 2. 학위논문 검색시스템 구성도

본문 DB를 검색함으로써 원하는 페이지만을 전송하여 인쇄할 수 있다.

5. 결 론

현대는 대량의 문서들이 다양한 형태로 생산되고 있다. 종이로 인쇄된 문서, 아래아한글, MS WORD등과 같은 다양한 종류의 문서편집기로 작성된 문서, 그리고 음성이나 동영상등 멀티미디어 정보등 수를 셀 수 없을 정도로 넘쳐나고 있다. 이러한 정보들을 사용하기 위하여 국내에서는 전자도서관 구축에 온 힘을 쏟고 있다. 그러나 전자도서관의 기술 미흡과 서비스 문제가 논란의 대상이 되고 있으며 관심과 노력에 비해 기술적인 측면에서는 아직도 열악한 상황이라 할 수 있다.

본 연구에서는 전자형태 2,236건에 대한 전자문서DB와 텍스트전문DB을 구축하였고 이를 연구개발정보센터의 검색시스템인 KRISTAL-II를 통해 사용자들에게 서비스하고 있다. 그리고 전자형태 원문에 대한 페이지 단위 검색 및 전송기능등을 갖추게 됨으로써 사용자들에게 편리성을 제공함과 아울러 향후 전자형태 문서에 대한 DB화의 길을 여는데 크게 기여하리라 생각된다.

6. 참고문헌

[1] 연구개발정보센터, "전자도서관 인프라 및 DB구축", 1998

[2] 연구개발정보센터, "정보검색을 위한 효율적인 저장 시스템 개발", 1997

[3] 연구개발정보센터, "인터넷 인트라넷 환경에서의 온라인 문서관리를 위한 MS-Word형식문서의 처리에 관한 연구, 1998

[4] 연구개발정보센터, "인터넷을 통한 복합문서의 전송 및 처리방안에 관한 연구", 1998

[5] "DVI 문서형식과 PDF 문서형식의 비교", http://www.texplus.com/texplus/comp5.html

[6] 연구개발정보센터,"국내 과학기술분야 석박사 DB", http://www.kordic.re.kr/~thesis