학위논문 전문데이터베이스 구축 및 서비스환경 구현

이 기 호[†] · 김 진 숙[††] · 윤 화 묵[†††]

요 약

1990년대 중반부터 다양하고 강력한 문서편집기의 보편화와 더불어 국내외의 대학에서는 책자형태의 논문제출과 동시에 전자형태 학위논문의 제출을 의무화하고 있다. 그러나 제출된 방대한 양의 전자형태의 논문들은 아래아한글, MS-Word, LaTex 등 다양한 문서편집기로 작성되었고 문서형식의 표준화가 이루어지지 않아 효율적으로 활용되지 못하고 있는 실정이다. 본 논문에서는 다양한 형태로 존재하는 학위논문들을 하나의 통일된 중간포맷으로 변환하고, 변환된 논문들을 전문데이터베이스(Full Text Datsbase)화하여 이를 인터넷을 통해 효율적으로 검색하고 서비스하기 위한 학위논문 전문검색시스템을 구현한다.

Construction of Full-Text Database and Implementation of Service Environment for Electronic Theses and Dissertations

Kyi-Ho Lee[†] · Jin-Suk Kim[††] · Wha-Muk Yoon[†††]

ABSTRACT

From the middle of 1990s, most universities in Korea have requested their students to submit not only the original text books but also their Electronic Theses and Dissertations(ETD) for masters degree and doctorates degree. The ETD submitted by the students are usually developed by various kinds of word processors such as MS-Word, LaTex , and HWP. Since there is no standard format for ETD to merge various different formats yet, it is difficult to construct the integrated database that provides full-text service. In this paper, we transform three different ETD formats into a unified one, construct a full-text database, and implement the full-text retrieval system for effective search in the Internet environment.

1. 서 론

1990년대 중반이래 웹을 기반으로한 정보통신기술의 급격한 발달과 고성능 검색엔진의 출현은 전세계가 다양한 분야의 정보를 시간과 공간을 넘어 공유할 수 있게 하고 있다. 또한 정보의 형태도 전통적인 텍스트위주의 정형적인 문헌정보로부터 비디오, 오디오, 이미지 등의 복잡한 멀티미디어 형태까지 다양하게 구성되고 있다. 특히 문헌정보의 경우 지금까지의 책자형태뿐만 아니라 전자형태로 작성되거나 OCR기법에 의해 변환됨으로써, 다양한 서버에 저장되어 이를 인터넷상에서 온라인으로 접근하고 원문을 검색할 수 있는 서비스체제가 일반화되고 있다. 많은 저명한 학회지들이 논문을 온라인으로 접수, 선정, 편집, 출판하는 전자저널(Electronic Journal)시스템을 도입하여 실제적으로 운영하는 것이 점점 보편화되고 있는 추세이다.

이러한 문서들의 전자 형태화에 의한 중요한 변화는 전통적인 서지 데이터베이스를 중심으로한 검색방식에서 전문데이터베이스(Full Text Database)를 대상으로

† 정 회 원 : 연구개발정보센터 선임연구원
†† 준 회 원 : 한국과학기술원 대학원 전산학과
††† 정 회 원 : 연구개발정보센터 연구원
논문접수 : 1999년 8월 20일, 심사완료 : 1999년 11월 8일

하는 검색방식으로의 변화이다. 간단하고 요약된 서지정보보다는 온라인으로 원문을 추구하는 사용자요구를 충족하기 위해, 90년대 이래 온라인 상용 시스템에는 전문데이터베이스가 계속해 증가하여 왔고, 최근에는 전문데이터베이스가 차지하는 비율의 증가는 주목할 만하다[7]. 전문데이터베이스는 "문서에 포함되어 있는 모든 문헌들의 문장, 단어, 그리고 모든 문자들을 컴퓨터에 저장시킴으로써 기계에 의해 이들을 검색할 수 있도록 만들어진 데이터베이스로 정의"되고 있다[10].

인터넷환경에서의 자동전문검색시스템은 많은 이점을 제공하는데 Blair와 Maron은 아래와 같이 설명하고 있다. 첫째로, 정보 및 컴퓨터 기술의 끊임없는 발전은 보다 빠르고 저렴하며, 신뢰도가 높고 사용하기 편리한 컴퓨터로의 접근을 가능하게 하고 있으며, 둘째로, 고용료가 비싸고 때로는 색인작업의 일관성을 유지하지 못하는 색인자가 필요 없고, 셋째로 정보원인 전문데이터는 출판을 위해 원고를 준비하는 과정에서 쉽게 얻어질 수 있으며 마지막으로, 문헌의 전문을 탐색함으로써 보다 정확하게 탐색어에 접근할 수 있다[11].

다양한 문서편집기의 보편화에 따라 국내 대학에서도 몇년 전부터 석박사 학위논문을 책자와 더불어 전자형태로 제출하도록 하는 곳이 점점 늘어나고 있다. 1997년부터 KAIST와 포항공대는 전자형태의 학위논문의 제출을 제도화했으며, 최근에는 국립대학을 중심으로 많은 대학들이 참여하고 있다. 그러나 제출된 전자형태의 학위 논문들은 아래아한글, MS-Word 및 TeX 등과 같이 각양각색의 문서편집기로 작성되었고, 또한 문서형식이나 스타일 등의 표준화 부재로, 대부분의 학위논문들이 인터넷을 통하여 효율적으로 제공되고 있지 않는 실정이다.

본 논문의 주요 목적은 현재 활용되지 못하고 있는 각 대학들의 방대한 양의 전자형태의 논문들을 하나의 전문데이터베이스로 구축하고 이를 활용하는 것이다. 첫째, 여러 형태의 전자문서를 하나의 데이터베이스로 통일시키기 위한 효율적인 복합문서 변환 및 처리방법을 제시한다. 둘째, 구축된 전문 데이터베이스를 인터넷상에서 효율적으로 서비스할 수 있는 전문검색시스템을 구현한다.

본 논문의 구성은 다음과 같다. 제2장은 다양한 형태의 전자문서들의 특징과 차이를 비교하고 문서의 변환과 전송의 기준을 제시한다. 중간포맷으로 DVI를 선정한 배경을 설명하고 문서변환을 위해 개발한 각종 도구들을 기술한다. 제3장에서는 정보검색시스템으로 사용한 KRISTAL-II를 간략히 소개하고 전문데이터베이스 구축 과정을 설명한다. 전문검색시스템 구성도를 통해 시스템의 기본 구조와 전문검색과정을 단계별로 보여준다. 제4장에서는 결론을 제시한다.

2. 문서의 변환 및 처리

2.1 배 경

1990년대 중반 이래 미국과 유럽 대학들은 학위논문 제출시 논문작성의 형태를 전통적인 책자형태에서 전자문서 형태로의 변환을 시도하기 시작하였다. 이러한 경향은 MS-Word, 워드퍼펙트, LaTex 등 다양한 기능을 보유한 강력한 워드프로세서의 일반화로 학생들이 경제적으로 전자문서를 쉽게 제작할 수 있기 때문에 가능하게 되었으며, 점차 많은 대학들은 이러한 전자형태의 학위논문제출을 제도화하려고 노력하고 있다. 우리나라에서도 1997년부터 한국과학기술원과 포항공대를 중심으로 책자형태의 논문과 동시에 전자문서 형태의 논문을 제출하도록 하고 있으며, 현재는 다른 대학들도 전자형태의 논문제출을 의무화하려는 추세이다.

그러나 각각의 워드프로세서에 의해 작성된 학위논문들을 활용하는데 가장 큰 문제점은 원본 파일의 다양한 내용과 편집상태를 그대로 유지하면서 인터넷에서 서비스할 수 있는 효율적 방법이 없다는 것이다. 지금까지 학위논문같이 복잡한 문서와 관련해 인터넷에서 사용할 수 있는 프로그램들은 단순히 문서를 HTML로 변환하는 것이 대부분이다. 예를 들면 아래아한글, MS-Word, 워드퍼펙트 등과 같은 워드프로세서들은 이들로 작성한 문서를 HTML로 변환하여 주는 도구프로그램이나 CGI, 플러그인 등을 개발하여 제공하고 있는 정도이다.

학위논문 같이 복잡한 문서들은 가장 일반적인 문서에서 사용하는 MS-Word로부터, 수식이나 수학기호 표현이 용이한 LaTex와 Mac, 한글의 경우 아래아한글이나 훈민정음 등까지 여러 가지 전자문서 형태로 구성되어 있다. 그러므로 다양한 복합문서를 하나의 데이터베이스에 저장하고 전문검색기능을 부여하기 위해서는 하나의 표준화된 파일 형태를 결정한 후, 모든 문서형태를 표준포맷으로 변환시켜야 하며, 그 다음 새로 생성된 문서들로부터 전문검색이 가능하도록 하기 위한 텍스트추출, 자동색인의 과정 등의 과정을 수

행하여야 한다.

2.2 형태 변환 및 전송의 기준

다양한 형태로 작성된 방대한 양의 학위논문들을 자동으로 전문데이터베이스화하기 위해서는 표준형태로의 효율적인 문서 변환이 이루어져야 한다. 즉 아래아한글, MS-Word, LaTex 등 여러 가지 전자형태의 학위논문에 대해 원본 파일의 다양한 내용과 편집상태를 그대로 유지하면서 인터넷에서 원문과 동일하게 서비스할 수 있는 문서변환의 원칙을 마련하여야 한다. 또한 복잡하고 용량이 큰 문서를 인터넷에서 신속하게 전송하기 위한 문서분할 같은 기법도 고려하여야만 한다. 이러한 효율적인 문서변환 및 전송을 위해 본 논문에서는 복합문서 변환기준을 다음과 같이 설정하였다.

- 원문에 대한 전문검색이 가능해야 한다.
  원문으로부터 페이지 단위의 검색을 할 수 있도록 하여 검색의 효율성을 높이도록 한다.
- 텍스트 정보를 추출할 수 있어야 한다.
  원문에 대한 전문검색이 가능하도록 하기 위해서는 변환된 파일형식으로부터 정보검색시스템의 색인 및 검색을 위한 텍스트 정보를 추출할 수 있어야 한다.
- 문서의 크기가 작아야 한다.
  일반적으로 복합문서는 문서의 크기가 인터넷상에서 전송하기에는 너무 크기 때문에 적당하지 않다. 변환하고자하는 파일 형식의 크기는 인터넷에서 전송하기 적당한 크기를 유지해야 한다.
- 원문의 페이지 단위 전송이 가능해야 한다.
  사용자들은 검색한 해당 페이지만을 보기 원하므로 하나의 문서에서 원하는 페이지를 추출하여 전송 가능해야 한다.
- 전송도중 열람이 가능해야 한다.
  문서 전체를 보는 경우 우선 전송된 페이지들만 먼저 보여 줄 수 있어야 한다.
- 문서 형태간의 호환성이 있어야 한다.
  구축된 파일 형식으로부터 향후 타 문서형식(PDF, PS, DVI 등)으로 변환할 수 있어야 한다.

2.3 표준 문서 형태

2.3.1 전자문서형식의 비교

급변하는 정보통신분야의 특성 때문에 인터넷 서비스를 위한 전자문서의 표준형태를 결정한다는 것은 가장 어려운 일 중의 하나이다. <표 1>은 지금까지 가장 일반적으로 사용되는 인터넷 전자문서의 대표적 형식들에 관한 비교를 나타내고 있다. 이 표는 각 형식에 대한 특징과 개요를 설명하고, 각 생성프로그램 및 뷰어에 대한 소개와 각 형식의 상대적 크기를 비교해서 보여준다[6].

현재 웹에서 가장 많이 사용하는 인터넷 전자문서형식은 미국 Adobe사의 PDF라 할 수 있다. 아래아한글이나 MS-Word 등과 같은 워드프로세서들은 단순히 문서를 HTML로 변환하여 주는 도구 프로그램이나 CGI, 플러그인을 개발하여 인터넷상에서 이들 문서를 처리하고 있으며 사실상의 표준 전자형식은 없다. 국제표준으로써 SGML이 제안되어 시험적으로 사용되고 있으나 복잡성과 제작비용의 과다 등의 문제점이 야기되어 아직은 실용화되지 못하고 있다. 최근에는 SGML의 대안으로 XML이 제안되어 업계를 중심으로 시험 DTD와 관련 시스템을 개발 중에 있으나 그 미래는 미지수인 실정이다.

<표 1> 전자문서형식 비교

(크기는 DVI가 1일 경우 상대적 크기)

형 식	개 요	생성프로그램	열람프로그램	크기
DVI	DeVice Independent 형식	TeXplus (HWP) Write, TeX	TeXplus Viewer, 각종 DVI Viewer	1
PS	PS 프린터와 그래픽을 위한 형식	PS프린터 드라이버, DVIPS	GhostScript/ Viewer	2 영문
PDF	Adobe가 만든 압축 PS형식	AdobePDF Write, Distiller, DocuCom Distiller	Acrobat Reader, DocuCom Viewer	5
DOC	MS Word 문서	MS Word	MS Word Viewer	6
HWP	아래아 한글 문서	아래아 한글	나모 Viewer	4
XLX	레이저젯 프린터를 위한 PCL에 기반	JetDoc	JetDoc Viewer	3
XML	SGML의 실현 가능한 형식	MSOffice 99:합의된 표준여부와 한글 지원은 미지수	Internet Explorer, Netscape(?)	?

국내 학위논문의 경우 아래아한글, MS-Word, PDF, SGML, DVI, PS 등 많은 형태로 구성되어 있기 때문

에, 이중에서 무엇을 선택하느냐는 단순한 결정이 될 수 없으며 각 형태로의 변환에 따른 기술적 문제, 경제성, 비영어권 문자여부, 파일크기 등 많은 요인들에 의해 결정되어야 한다. 특히 국내 석박사 논문의 경우 대부분이 한글과 영어로 구성되었기 때문에 한글의 특성이 고려되어야만 한다. 즉, 한글 문서에 전문 검색 기능을 부여하기 위해서는 한글폰트에 대한 지원여부와 파일크기, 색인어 추출문제, 문서 형태간의 호환성 여부, 향후 업계의 지원계획 등의 문제해결이 선행되어야 한다.

2.3.2 PDF와 DVI

미국의 학위논문 데이터베이스 구축의 경우, 미국정부의 지원하에 1990년대 중반부터 버지니아공대(Virginia Tech)와 UMI(University Microfilms Inc.)를 중심으로 전자형태의 학위논문들을 수집하여 데이터베이스를 구축하고 있는데 PDF를 기본형식으로 채택하고 있다[8]. 학생들은 MS-Word, 워드퍼펙트, LaTex, SGML 4가지 형식의 전자형태 학위논문을 제출할 수 있는데, 이 논문들이 데이터베이스화 될 때는 PDF로 변환되어 저장된다. SGML형식의 문서는 자체적으로 DTD제작, 문서편집기 및 뷰어시스템 등을 개발하여 시범적으로 관리하고 있는 정도이다[9].

한글문서의 경우 PDF를 채택하는데는 다음의 문제점들이 발생한다. 먼저 PDF는 한글과 같은 비영어권의 2바이트 문자표현이 고려되지 않았다는 단점이 있다. 가장 큰 문제는 PDF의 경우 한글을 처리하기 위해서는 한글 서체를 포스트스크립트 서체로 변환하여 내부에 담기 때문에 파일의 크기가 수 메가바이트에 달하게 되어 인터넷을 통한 문서의 전송 및 서비스가 부적합하다. 또한 전문데이터베이스 구축에 절대적으로 필요한 기능인 검색어 추출에 의한 전문검색을 한글 PDF문서에서는 지원하지 않는다. 물론 포스트스크립트 파일로부터 Distiller를 통해 PDF를 생성할 수 있지만 한글부분은 문자코드값이 바뀐 일종의 그림과 같이 저장되어 전자문서로서의 기능을 저하시킨다.

본 연구에서는 학위논문의 표준문서형식으로 DVI를 채택하였다. 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. 또한 DVI에서 PDF로의 변환은 간단한 소프트웨어로 가능하기 때문에, 위에서 언급한 한글에서의 PDF문제가 해결되면 언제든지 PDF형식으로의 호환이 가능하다. <표 2>는 PDF와 DVI 형식을 비교한 것인데, 한글문서의 전자 형태화에는 상당히 효과적으로 나타나 있다[6].

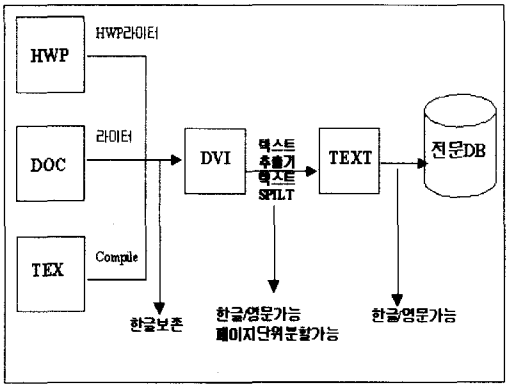
〈표 2〉 DVI와 PDF의 비교

	DVI 형식	PDF 형식
글 꼴	Windows 형식인 TrueType	Adobe 형식인 PS Type1과 4
글 꼴 처 리	내장하지않고 시스템 및 자체 글꼴로 대체	영문글꼴은 자체글꼴로 대체하고 한글글꼴은 문서에 내장
삽입된 벡 터 그래픽	문서에 압축내장, EPS와 Windows 형식인 WMF동시지원	문서에 압축 내장, EPS기반
삽입된 비트맵 그래픽	종류에 따라 JPG 및 GIF압축내장	종류에 따라 JPG 및 GIF압축내장
텍스트	공개된 표준 압축방식, 색인과 검색엔진 선택이 자유로움	Adobe 압축방식, 색인과 검색엔진에 제약이 따름
수 식	TeX와 MathType 수식을 위해 자체글꼴 사용	수식서체 문서에 내장

2.4 문서의 변환

지금까지 국내에서 생산된 전자형태 석박사 학위논문들의 파일형식을 분석해 보면 아래아한글, MS-Word, LaTex 등 3종의 문서형식이 거의 대부분으로 약 90% 이상을 차지하고 있다[5]. 본 연구에서는 각기 다른 형식(HWP, DOC, TEX)으로 저장되어 있는 전자문서를 동일한 형식(DVI형식)으로 변환하기 위한 DVI 처리도구를 개발하였다[4]. 문서변환기로써 TeXplus 라이터, TeXplus HWP 라이터를 개발하였으며, 페이지 단위로 문서를 검색하기 위하여 DVI형식의 문서로부터 텍스트를 추출하는 텍스트 추출기(DVI2TXT), 그리고 DVI 문서에서 특정 페이지를 추출하기 위한 DVI서버(DVI Split)와 DVI문서를 브라우저로 볼 수 있는 DVI뷰어(TeXplus)를 개발하였다. 개발된 DVI 처리 도구들의

구성도는 (그림 1)에 나타나 있으며, 각 변환기와 처리기의 주요 특징과 기능은 다음과 같다.



(그림 1) 문서변환 및 문서적재

2.4.1 TeXplus 라이터

TeXplus 라이터는 마이크로소프트 윈도우 95/98/NT상에서 한글과 컴퓨터사의 아래아한글을 제외한 응용 프로그램에서 DVI화일을 생성시켜주는 프린터드라이브 소프트웨어다. 특히 우리나라에서 많이 사용하고 있는 한글 워드프로세서인 ㈜마이크로소프트의 한글워드, 삼성전자의 훈민정음, 핸디소프트의 아리랑 등에서 원본과 동일하게 수식과 그래픽을 포함한 한글 DVI파일을 만들수 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].

2.4.2 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[1].

2.4.3 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.

2.4.4 DVI 본문추출기

TeX의 중요한 실행 결과물인 DVI 파일의 내용을 보려면 xdvi 같은 DVI 뷰어가 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 한글로된 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 텍

스트로 추출한다. 이렇게 추출된 텍스트는 정보검색시스템에 적재하여 문서 데이터베이스로 저장할 수 있다. 사용자들은 이 데이터베이스를 검색함으로써 DVI 문서를 페이지 단위로 검색할 수 있게 된다.

2.4.5 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.

3. 전문검색서비스시스템 구현

3.1 KRISTAL-II

KRISTAL-II 시스템은 연구개발정보센터에서 과학기술정보서비스의 기본 시스템으로 사용하기 위해 독자적으로 개발된 시스템으로, 한글과 한자 및 영문이 혼용된 문서를 효율적으로 저장하고 검색할 수 있는 정보검색시스템이다. KRISTAL-II는 현재 연구개발정보센터의 과학기술서비스의 기본 시스템으로 활용되고 있을 뿐만 아니라 정보통신부 주관으로 추진 중인 국가 주요 전자도서관 사업의 기본시스템으로 활용되고 있으며 그 외에도 표준연구소 등 20여개의 연구기관에서도 사용되고 있다.

KRISTAL-II는 서지정보검색에 적합하도록 가변 길이의 필드를 여러 개 가지는 문서를 검색 대상으로 하고 있으며, 절단연산자와 근접도 연산자 그리고 히스토리 검색기능을 갖추어 상용 검색 엔진에서 제공하는 기본적인 기능들을 모두 제공하고 있다. 이 시스템의 특징은 한글자동색인시스템이 추가되어 한글 문서를 효과적으로 처리할 수 있게 되었으며, 최대 문서 크기를 1.3M 바이트로 확장하여 멀티미디어 정보 저장을 위한 기본 저장 구조를 갖추었고, 다중 데이터베이스 검색이 가능하도록 구현되었다[2].

1999년에 발표된 KRISTAL-II Version 2.0은 전자도서관 구축에 필요한 몇가지의 기능을 추가하고 있다. 정보 검색에 익숙하지 않는 사용자도 KRISTAL-II를 쉽게 이용할 수 있도록 검색 결과의 랭킹 기능을 도입하고, KRISTAL-II 적재 기능을 개선하여 일반 적재 뿐만 아니라 자동으로 실시간 정보 수정/삭제/추가가 가능하도록 했다. 또한 상용 데이터베이스 시스템에 자료가 이미 적재되어 있을 경우에도 KRISTAL-II를 이용하여 검색이 가능하도록 KRISTAL-II 시스템과 상용 데이터베이스 시스템의 연동 기능을 부가했다[5].

3.2 전문데이터베이스구축

국내 대학에서도 95년부터 KAIST를 시작으로 석박사 학위논문을 책자와 더불어 전자형태로 제출하도록 하고 있으며, 최근에는 국립 대학들도 도서관의 전자화를 위해 이를 제도화하고 있는 추세이다. 그러나 각 대학들에 제출된 많은 양의 논문들은 대다수 디스켓으로 보관되어 있을 뿐 활용되지 못하고 있으며, 현재까지 논문들을 전문 데이터베이스화하여 인터넷을 통하여 효율적으로 서비스하고 있는 곳은 세계적으로도 거의 없는 실정이다. 본 연구에서는 활용되지 않고 디스켓형태로 보관된 대용량의 다양한 전자형태의 석박사 학위논문들을 중간 형태(DVI)로 변환한 후 전문검색이 가능한 데이터베이스를 구축하고 이를 서비스하는 시스템을 구현하였다

전문 데이터베이스 구축 대상은 KAIST와 포항공대, 두 대학으로 선정하였는데, 각 대학이 소장하고 있는 전자형태의 학위논문의 분포는 <표 3>과 같다. 전자형태 학위논문의 분포를 문서편집기별로 분류해 보면 DOC형식이 약 40%로 가장 많고, 다음 HWP, DVI의 순서로 나타내고 있다.

학위논문 문서의 적재는 학위논문 서지사항 데이터베이스와 전문 데이터베이스로 분리되어 KRISTALL-II에서 이루어진다. 서지사항은 각 대학이 보유하고 있는 서지정보를 연구개발정보센터에서 표준형식으로 통합하여 KRISTAL-II 데이터 관리기에 적재하고 학위논문을 검색하기 위한 기본 자료로 활용하고 있다.

〈표 3〉 문서편집기별 전자형태 학위논문 분포

대학별	HWP	DOC	DVI	계
KAIST	617 (32%)	863 (45%)	431 (23%)	1911
포항공대	110 (34%)	136 (42%)	79 (24%)	325
계	727 (33%)	999 (44%)	510 (23%)	2236

각각 다른 형식으로 저장되어 있는 전자문서들은 문

서변환기(Texplus라이터, Texplus HWP라이터)를 이용하여 DVI형태로 변환하였다. 이렇게 생성한 DVI파일은 전문검색을 위하여 KRISTAL-II DVI서버에 보관하게 된다. DVI파일은 서버의 적당한 위치에 저장하고, 색인파일로 재구성하여 저장하게 된다. 페이지 단위로 추출된 텍스트정보는 문서 적재를 위하여 필요한 스키마를 작성함으로써 KRISTAL-II에 적재한다.

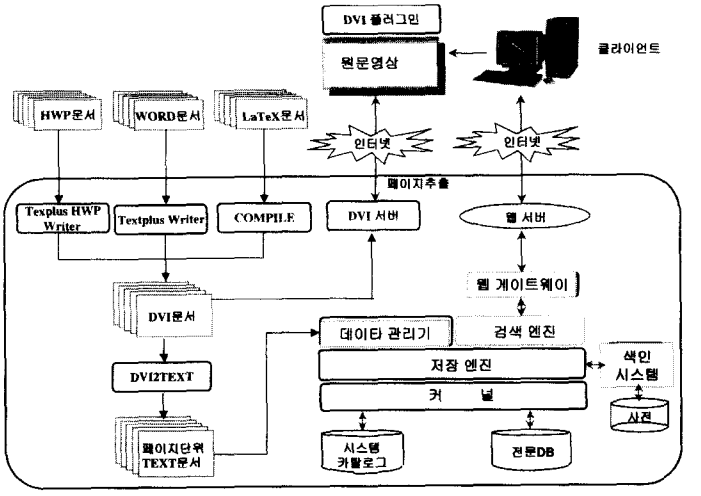
3.3 검색시스템 구성도

검색시스템의 기본적인 구조는 사용자가 학위논문 서지정보 데이터베이스를 검색함으로써 각 학위논문의 초록을 먼저 검색하고 이를 통해 DVI 형태의 원문을 검색하는 2단계 검색과정으로 구성되었다. 인터넷에서 검색속도의 향상을 위하여 본문 데이터베이스를 검색할 때에는 원하는 페이지만을 전송하여 검색하고 인쇄하도록 DVI 문서분할기를 사용하였다. 특히 자연어 질의를 통해 텍스트전문에 대한 검색 뿐만 아니라 그림이나 표도 검색하기 위해 그림과 표에 있는 텍스트로부터 색인어를 추출하여 검색시 활용할 수 있도록 디자인하였다.

(그림 2)는 학위논문 검색시스템 구성도를 나타내고 있다. 전문검색은 사용자가 자연어 질의를 사용하여 페이지단위로 검색하는데 전문검색의 단계는 다음과 같다.

1) 사용자들은 클라이언트에서 검색할 질의어를 입력한다.
2) 웹 게이트웨이는 KRISTAL-II 검색엔진에 적합한 질의어로 변환한다. 검색엔진은 적재되어 있는 전문 데이터베이스로부터 검색을 수행하여 그 결과를 웹게이트웨이를 통해 클라이언트에 보내준다.
3) 페이지 단위로 검색된 화면에서 사용자는 자신이 원하는 원문을 보기 위해 해당 페이지 요약문을 선택한다. 단, 제목을 선택하면 처음 페이지부터 볼 수 있다.
4) 웹 게이트웨이는 페이지번호, 이미지 파일명, 문서위치정보를 추출한다.
5) 웹 게이트웨이는 문서위치 정보를 이용하여 DVI 이미지를 추출하여 해당 페이지를 가져오도록 DVI서버를 호출한다. DVI서버는 페이지번호를 이용하여 해당 페이지만을 추출하고 DVI 뷰어로 제공한다.
6) DVI 뷰어는 해당 DVI이미지를 화면에 출력한다. 이때 DVI 뷰어는 다음과 이전페이지, 특정페이지 보기 기능을 통해 원문을 계속적으로 볼 수 있다.

서비스 시스템 구성도를 보면 전문검색 데이터베이스뿐만 아니라 이미지형태 데이터베이스 서비스도 제공하고 있다. 1990년 중반 이전의 학위논문은 거의 책자형태로 보관하고 있으므로 이들에 대해서는 TIFF형식의 이미지형태로 데이터베이스를 구축하였다. 사용자들은 학위논문의 서지사항을 검색하여 자신이 원하는 원문을 이미지로써 접근할 수 있고, 또한 TOC (Table of



(그림 2) 학위논문 검색시스템 구성도

Contents)를 통해 원하는 페이지의 원문을 직접 열람하도록 디자인하였다.

학위논문에 기술된 기본목차와 표목차 그리고 그림목차, 즉 문서구조정보를 통해 바로 원문에 접근을 허용하는 것은 사용자에게 상당한 편의를 제공한다. 문서구조정보는 국내 전자도서관사업 등 타기관과의 문서교환을 감안하여 SGML로 제작하였다 TOC는 TIFF 형식이 아닌 타 매체를 수용할 수 있도록 구성하였으며, TIFF 이미지를 PDF 등 타 이미지로 변환했을 때도 사용할 수 있도록 독립적으로 저장하고 관리토록 하였다.

3.4 저작권문제

전자형태로 제출된 학위논문을 전문 데이터베이스화하여 서비스하기 위해서는 다른 창조적 디지털 저작물과 같이 저작권문제 이슈가 해소되어야 한다. 논문의 저자는 자신의 학위논문 이용에 관해 배타적권리(Exclusive Right)를 법적으로 부여받고 있으므로, 구축된 데이터베이스를 공개하기 위해서는 저작권자의 동의를 얻는 것이 필수적이다[12].

우리나라에서는 KAIST, 포항공대, 광주과학기술원 등의 대학들이 1990년 중반부터 전자형태의 학위논문 제출의 제도화를 시도하면서, 논문 제출시 공개동의서 제출을 의무화하였다. 공개동의는 전문공개, 부분공개, 조건부공개(1년후 공개), 비공개 등으로 구분하였으며, 전문 검색시스템은 이러한 구분 등급에 따라 제한적으로 공개하도록 구현하였다. 본 연구에서 구축한 KAIST와 포항공대의 전자형태 학위논문의 공개동의 현황은 <표 4>와 같다. 표를 보면 약 95% 이상의 학생들이 공개에 동의하고 있는데, 산업계와의 지적재산권이 관련된 논문들을 포함한 일부를 제외하고는 대부분이 논문의 공개에 찬성하고 있다.

〈표 4〉 전자형태 학위논문 공개동의 현황

대학별	공개 동의	비공개	미확보	공개 동의율
KAIST (1995-1997)	1816	38	57	95%
포항공대 (1999)	325	0	0	100%

그러나 1990년 중반 이전의 책자형태의 학위논문은 저작권 동의가 거의 없는 상태이므로, 이를 이미지 기반으로 데이터베이스화하여 공개하기 위해서는 저작권 해소가 선행되어야 한다. KAIST와 포항공대 두 대학은 '97년 이전의 책자형태 저작권 해결을 위해 1998년부터 우편, 전자우편, 신문광고, 동창회 등을 통해 각 논문저자와 접촉해 저작권 공개동의서를 얻고 있으며 이를 현재까지 지속적으로 추진하고 있다.

1997년 이전 학위논문 공개동의 현황을 조사한 결과는 각 대학별로 평균 40%정도의 동의를 보이고 있으며, 비공개의 경우 대부분 저자와의 연락 불가 때문에 공개 동의를 얻지 못하고 있다[1]. 이러한 연락 단절에 의한 비공개의 경우는 현재 저작권법상으로 해결할 방법이 없으며, 전체논문의 절반을 얻는다는 점에서 앞으로 해결해야 할 중요한 이슈로 나타내고 있다.

4. 결 론

현대는 대량의 문서들이 다양한 형태로 생산되고 있다. 특히 문헌정보는 전통적인 책자형태 뿐만 아니라 다양한 문서편집기에 의해 전자형태로 제작되고 있다. 강력하고 다양한 문서편집기에 의한 문서들의 전자형태화에 따른 가장 큰 변화는, 지금까지의 서지데이터베이스를 중심으로한 검색방식에서 전문데이터베이스를 대상으로하는 검색방식으로의 변화이다. 인터넷의 일반화로 사용자들은 요약된 서지정보보다는 온라인으로 원문을 요구하게 되었고, 이에 따라 온라인 상용 시스템에는 전문데이터베이스가 계속해 증가하여 왔다.

1990년 중반부터 국내 대학에서는 석박사 학위논문을 책자와 함께 전자형태로 제출하도록 하는 곳이 점점 늘어나고 있으며, 그 양도 점차 증가하고 있다.

지금까지 각 대학들에 제출된 전자형태의 학위 논문들의 구성을 보면, 아래아한글, MS WORD 및 TeX 등과 같이 각양각색의 문서편집기로 작성되었고, 또한 문서형식이나 스타일 등의 표준화 부재로, 대부분의 학위논문들이 인터넷을 통하여 효율적으로 제공되고 있지 않는 실정이다.

본 논문에서는 활용되지 못하고 적재되어 있는 전자형태의 논문들을 자동으로 전문데이터베이스화하기 위한 효율적인 복합문서 변환방법과 인터넷상에서 효율적으로 서비스할 수 있는 전문 검색시스템을 개발하였다. 적용사례로써 KAIST와 포항공대의 전자형태논문 2,200건에 대한 텍스트전문데이터베이스를 구축하였고 이를 연구개발정보센터 정보검색시스템인 KRISTAL-II를 사용하여 서비스 환경을 구현하였다.

이러한 결과는 전자형태 원문에 대한 페이지 단위 검색 및 전송 기능 등을 갖추게 함으로써 사용자들에게 편리성을 제공함과 아울러 향후 전자형태 문서에 대한 데이터베이스화의 길을 여는데 크게 기여하리라 생각된다.

참 고 문 헌

[1] 이기호 외, "전자도서관 인프라 및 데이터베이스 구축", pp.263, 연구개발정보센터, 1998.

[2] 이준호 외, "정보검색을 위한 효율적인 저장시스템 개발", pp.162, 연구개발정보센터, 1997.

[3] 유성준 외, "인터넷/인트라넷 환경에서의 온라인 문서관리를 위한 MS-Word형식문서의 처리에 관한 연구", pp.22, 연구개발정보센터, 1998.

[4] 서영진 외, "인터넷을 통한 복합문서의 전송 및 처리 방안에 관한 연구", pp.83, 연구개발정보센터, 1998.

[5] 박혁로 외, "효율적 정보검색 환경구현", pp.203, 연구개발정보센터, 1998.

[6] DVI 문서형식과 PDF 문서형식의 비교, http://ww.w.texplus.com/texplus/comp5.html

[7] 문성빈, "적합성 피드백을 이용한 전문검색시스템의 검색 효율성 증진을 위한 연구", 정보관리학회지, 제10권, 제2호, pp.43-67, 1993.

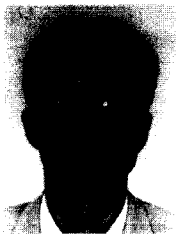
[8] Status of ETD Initiatives in the US and Canada, http://www.fis.utoronto.ca/etd/report1.html.

[9] An SGML/HTML Electronic Thesis and Dissertation Library, http://www.stg.brown.edu/webs/tei10/tei10.papers/erickson.html.

[10] Blair, D.C., & Maron, M.E., "Full Text Information Retrieval : Further Analysis and Clarification, Information Processing and Management," Vol.26, No.3, pp.437-447, 1990.

[11] Blair, D.C., & Maron, M.E., "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," Communication of the ACM, Vol.28, No.3, pp.289-299, 1985.

[12] Bennet & Scott, "Copyright and Innovation in Electronic Publishing : A commentary," Journal of Academic Librarianship, Vol.19, No.2, pp.87-91, 1993.



이 기 호

e-mail : ghlee@ng.kordic.re.kr

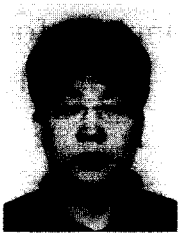
1973년 서울문리대 해양학과 (학사)

1985년 Bowling Green State Univ. 전산학과(전산학석사)

1999년 충남대학교 컴퓨터공학과 (공학박사)

1990년~현재 연구개발정보센터 선임연구원

관심분야 : 정보검색시스템, 전자도서관, 전자상거래



김 진 숙

e-mail : kjs@ns.kordic.re.kr

1993년 KAIST 생물공학과(학사)

1995년 KAIST 생명과학과(석사)

1995년~현재 연구개발정보센터 연구원

1998년~현재 KAIST 전산학과 (석사)

관심분야 : 정보검색, 인터넷정보서비스, 전자상거래

윤 화 묵

e-mail : hmyoon@ns.kordic.re.kr

1992년 서울산업대 전산학과 (학사)

1997년 공주대 전산학과(석사)

1990년~현재 연구개발정보센터 연구원

관심분야 : 원문서비스, 정보검색, 멀티미디어 정보서비스