복합 문서 개체 검색 시스템-[IN2] DOR

모비코앤시스메타(주) 부설 기술연구소
안태성 · 임중수 · 김명훈 · 안우람 · 이경일

Composite Document Object Retrieval and Searching System-[IN2] DOR

Tae Sung Ahn, Joong Su Yim, Myung Hoon Kim, Woo Ram Ahn, Kyung Il Lee
Mobico & Sysmeta Co. Ltd., Korea

요 약

기존 문서 검색 시스템의 경우 단순히 문서 내에서 텍스트를 추출한 후 그 텍스트를 색인, 검색하는 형태를 가지고 있었다. 본 논문에서는 MS Word, Excel, HWP 등 다양한 형태의 문서에서 텍스트, 표, 이미지, 차트, 동영상 등의 문서 개체를 분석, 색인하고 이를 검색하는 시스템의 개발 방법을 제안하였다. 제안된 시스템은 문서의 내부 자료 구조를 CDML(Composite Document Markup Language)로 변환하고, 이를 색인, 저장함으로 기존의 전문 검색 시스템의 한계를 효과적으로 극복 했으며, 문서 내의 검색 대상 개체로 자동 이동하고 하일라이팅 시키는 기술을 구현함으로 사용자 편익성을 높였다. 개발된 시스템의 성능을 평가한 결과, 다양한 문서 형식에 대해 평균 97% 이상의 CDML변환 성공률과 개체 검색 성공률을 보였으며, 이진 파일에서 직접 개체를 추출함으로 매우 높은 분석 및 색인 속도가 달성되었음을 확인할 수 있었다. 본 논문에서 소개된 새로운 패러다임의 문서 검색 솔루션을 통해 다양한 기술적 상업적 파급 효과가 기대되고 있다.

서 론

기업 환경이 점점 더 복잡 다양해지고, 인터넷 활성화를 통한 정보 유통량이 폭발적으로 증가됨에 따라, 보다 진보된 정보 검색에 대한 요구와 이에 따른 기술 개발이 활발히 진행되어 왔다. HTML 형식의 인터넷 문서를 위한 웹 문서 검색 시스템, HWP와 같은 다양한 형태의 문서를 대상으로 개발된 시스템, 동영상 등의 멀티미디어 콘텐츠를 검색하기 위한 시스템 등, 지난 10년 간 다양한 영역에서 검색 시스템이 개발, 발전되어 왔다. 최근 들어 인터넷을 활용한 학술 활동과 정보 교류가 활발해지면서, 모든 학술 문헌과 정보를 전자 문서 형태로 소장하고 온라인 상에서 활용하려는 디지털 도서관 구축이 적극 추진되고 있으며, 기업에서는 데이터베이스 중심의 정형화된 정보에 비해 워드, 스프레드시트, 프리젠테이션 파일 등의 다양한 비정형 문서가 기하급수적으로 증가함에 따라, 이들 디지털 자산의 관리와 재활용을 위한 EDMS(Electronic Document Management System) 및 KMS(Knowledge Management System) 솔루션의 도입이 활발히 추진 되고 있다. 이러한 디지털 자산 관리 시스템의 상업적 발전과 더불어 고객의 요구도 다양화, 전문화 되고 있으며, 기존의 전문검색(FTR) 시스템 뿐만 아니라 문서 내 고급 정보 검색, 보다 편리한 검색 결과 확인, 대량 문서에 대한 정보 분석 등의 기술적 요구도 크게 증가하고 있는 실정이다.

이러한 기술적 요구에도 불구하고 현재까지의 검색 시스템은 문서 내의 단순 키워드나 구문에 대한 텍스트 검색만을 지원하며, 텍스트 외에 표, 이미지, 차트, 그래프 등과 같이 보다 함축적인 '지식' 정보를 표현한 문서 개체에 대한 검색은 전혀 구현되어 있지 못한 실정이다.

본 논문에서는 단순한 텍스트 정보를 추출, 색인, 검색하는 기존 시스템의 한계를 극복하기 위해, 이진 형식의 복합 문서에서 텍스트를 포함한 다양한 문서 개체를 직접 추출 분석하고 이를 XML형식으로 변환, 색인하는 방법을 소개하였

E-mail : albert@mobico.com
E-mail : jslim@mobico.com
E-mail : hoital@mobico.com
E-mail : rambar@mobico.com
E-mail : tony@mobico.com

다. 또한, 고품질 다국어 형태소 분석 시스템에 기반한 전문 검색, 그래프, 이미지와 같은 문서 내 개체 검색, 자동 위치 이동 및 문장 하일라이팅 등의 획기적인 기능이 구현된 차세대 검색 시스템의 소개와, 그 성능 시험 결과를 고찰해 보고자 한다.

복합 문서(Composite Document)

복합 문서는 순수 텍스트 문서와 달리 다양한 속성이 부여된 텍스트 데이터와 차트 개체, 그래프 개체, 동영상 파일 등 서로 다른 형식의 데이터가 하나의 파일로 통합되어 있는 문서를 말한다. 이러한 복합 문서는 상호 연결이 가능하며, 사용자에 의해 실행 가능한 프로그램 개체가 포함되기도 한다.

1. 복합문서의 종류

현재 국내에서 가장 많이 사용되는 복합 문서는 한글과 컴퓨터사의 아래한글 파일(HWP), MS사의 오피스 문서(doc,xls,ppt), 그리고 어도비(Adobe)사의 PDF파일을 들 수 있다. 대부분의 복합 문서는 MS 오피스와 아래한글을 사용해 생성되고 있으며, Postscript에서 발전한 PDF 문서는 문서 배포의 목적으로 많이 사용되고 있다. 이 외에도 이러한 복합문서의 내용을 구조적으로 표현할 수 있도록 제안된 XML도 표준으로 널리 사용되고 있다.

1) 마이크로소프트 오피스

미국 마이크로소프트사에서 개발된 마이크로소프트 오피스는 워드, 엑셀, 파워포인트로 구성되어 있고, 오피스 2000 버전부터는 기존의 언어별 인코딩을 사용하지 않고, 유니코드 3.0을 완전히 지원함으로 자유로운 다국어 표현이 가능하게 되었다. 모든 오피스 문서들은 OLE Document라는 마이크로소프트사의 표준 복합 문서 저장 체제에 기반하고 있으며 이를 통해 버전별 최소한의 호환성과 표준화된 접근성을 유지하고 있다. 본 연구에서는 OLE Document 체제를 OS 비종속적으로 바이너리 차원에서 직접 접근하는 방법을 시도함으로, 매우 높은 수준의 복합문서 분석 및 변환 시스템을 구현하였다.

2) 아래한글

아래한글은 1989년 1.0 버전의 개발을 시작으로 현재 한글 2002 버전까지 개발되어 있다. 주로 학교와 관공서를 중심으로 많이 사용되고 있으며, 문서 포맷은 크게 815 버전 이전 형식과 한글 워디안 이후의 문서 형식으로 구별된다. 이 두 종류의 포맷은 상호 유사점 및 호환성을 가지고 있지 않으며, 별도의 분석 시스템을 필요로 하고 있다. HWP 형식의 문서를 바이너리 상태에서 분석하고 색인하기 위해 본 연구에서는 직접 문서 포맷 분석 파서를 제작하였다.

3) PDF

PDF 문서는 1990년 초 미국의 어도비사에서 기존의 PS(Postscript)를 개선하여, DTP 문서를 서로 교환할 수 있도록 개발하였으나, 인터넷의 폭발적인 증가와 다양한 OS에서 작동되는 전용뷰어의 무료보급으로 인해, 현재는 인터넷상의 문서 전달 목적으로 많이 사용되고 있다. 또한 오피스와 아래한글과 같은 기존의 문서편집기에서 PDF로 변환할 수 있는 기능을 제공함으로 더욱 더 활용이 늘어나고 있다.

복합 문서 개체 검색 시스템-[IN2] DOR

복합 문서의 개체를 수행하는 [IN2] DOR은 Fig. 2와 같이 기존 검색 시스템과는 다르게 색인 및 검색 모듈 외에 복합 문서를 CDML로 변환하는 문서 파서 모듈과 검색한 개체를 자동으로 선택하여 검색 활용성을 높여주는 개체 하이라이팅 모듈이 포함되어 있다.

1. 복합 문서 개체 변환

앞에서 소개한 복합 문서들에 들어 있는 다양한 개체에 대한 검색을 위해서는 적절한 형태의 문서 포맷 변환이 필요하다. 이를 위해 XML 기반의 CDML(Composite Document Markup Language)을 정의하였다.

Fig. 1에서는 CDML 문서의 예를 보여주고 있다. 각 복

```
<?xml version="1.0" encoding="UTF8"?>
<DORM-DOCUMENT-BODY>
<DORM-EXT>doc</DORM-EXT>
<DORM-PATH>민원처리부.doc</DORM-PATH>
<DORM-TEXT>[별지 제19호서식]</DORM-TEXT>
<DORM-TEXT>민 원 처 리 부(전화 및 상
담)</DORM-TEXT>
<DORM-TEXT>공사명 : OOO 건설공사</DORM-TEXT>
<DORM-CELL>월일</DORM-CELL>
<DORM-CELL>민 원 인</DORM-CELL>
<DORM-CELL>민 원 내 용</DORM-CELL>
<DORM-CELL>처리계획 및 조치내용</DORM-CELL>
<DORM-CELL>비 고</DORM-CELL>
<DORM-TEXT>( 주 소 )</DORM-TEXT>
<DORM-TEXT>( 성 명 )</DORM-TEXT>
<DORM-TEXT>( 연락처 )</DORM-TEXT>
<DORM-CELL>( 주민등록번호)</DORM-CELL>
<DORM-TEXT>( 위 치 )</DORM-TEXT>
<DORM-CELL>( 민원내용)</DORM-CELL>
<DORM-TEXT>주) 책임감리원이 처리 불가능한 사항은 발주
청에 서면 보고함</DORM-TEXT>
</DORM-DOCUMENT-BODY>
```

Fig. 1. CDML 문서의 예.

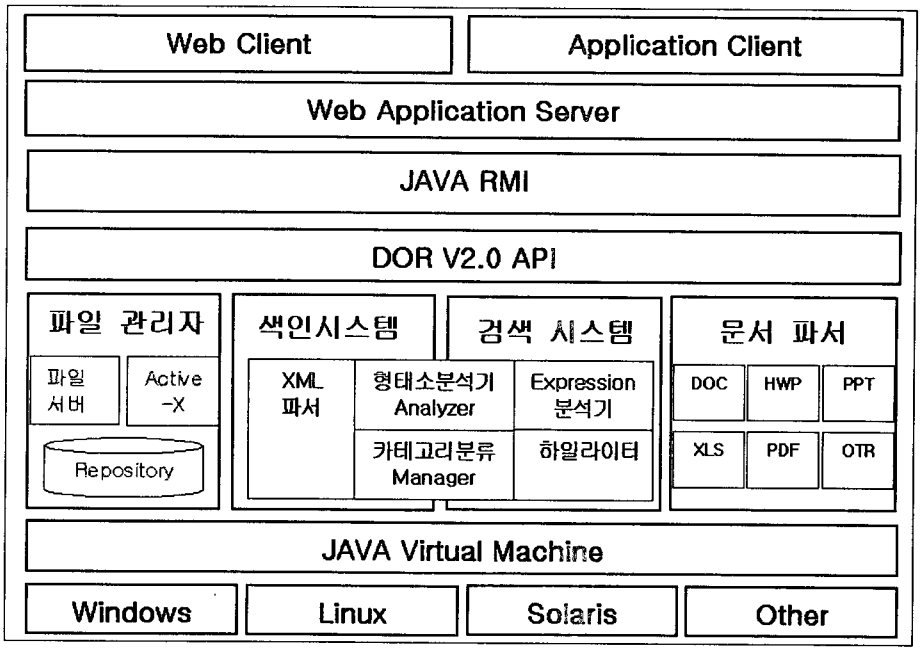

Fig. 2. [IN2] DOR의 시스템 구조.

합 문서 내의 텍스트 박스, 차트, 이미지 등을 모두 XML 형식의 CDML로 변환하기 위해서 본 연구에서는 바이너리 차원에서 직접 문서 분석을 시도하였다. 각각의 편집 소프트웨어 개발사에서 나름대로의 개발 라이브러리를 제공하고 있으나, OS 제한 및 속도 등의 문제로 각 문서의 파일 포맷을 직접 바이너리 차원에서의 접근하는 자체 문서 분석기 개발을 수행하였다. 기존의 텍스트 추출만을 지원하는 단순한 필터와는 달리 문서 내의 개체와 그 위치를 추출하기 위해서는 텍스트 영역뿐만 아니라 표, 그래프, 차트 등과 같은 개체 영역을 분석하여 각 개체형식과 데이터를 CDML 형식으로 변환하게 된다. 비 텍스트 형 개체들의 검색을 위해, 표와 같은 개체는 내부의 텍스트를, 이미지와 같은 개체는 주변의 텍스트를 사용해 색인을 수행한다. CDML에서 사용하는 인코딩은 다국어 처리가 가능하도록 하기 위해 UTF8을 사용했다. 변환 절차는 Fig. 3과 같다. 우선 처리할 복합 문서를 받아, 문서의 포맷과 버전을 확인한다. 이 때 문서의 확장자와 문서 포맷이 다른 경우에는 처리 가능한 문서 포맷으로 바꾸어 처리를 시도한다. 문서 포맷 판단이 된 후에는 문서 내용의 일부를 분석하여 사용된 언어를 판단한다. CDML 변환을 위한 각각의 문서 처리 모듈은 플러그인 방식으로 삽입할 수 있도록 설계하였기 때문에, 추후 새로운 처리 모듈만 추가하면 더 다양한 문서 형식에 대해서도 분석이 가능할 것이다.

문서 처리 모듈은 문서 분석 및 데이터 추출 프로세스와 분석 결과를 CDML로 변환하는 변환 프로세스로 구성되어 있다. 변환 프로세스는 XML 파서를 이용하여 CDML을 생

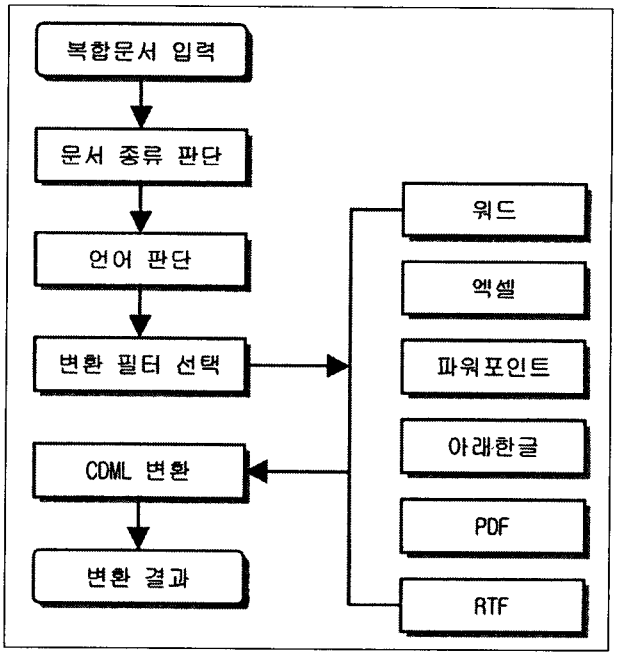

Fig. 3. 복합문서의 CDML 변환 과정.

성 및 분석하며, 색인기, 질의 분석기 등의 여러 모듈에서 공통적으로 사용되고 있다. 또한 CDML 변환 과정에서 각 언어별 인코딩을 UTF8 형식의 유니코드로 일원화했기 때문에, 세계의 모든 언어에 대한 색인 및 검색이 가능하도록 설계 되었다. 현재 개발 완료된 시스템에는 워드, 엑셀, 파워포인트, 아래한글, PDF에 대한 변환 필터들이 탑재되어 있다.

2. 색인 및 검색 시스템

색인 시스템은 개선된 역파일 기법이 사용되었다. 문서의 모든 부분을 색인 하는 것이 아니라, 고품질 형태소 분석을 통해 조사 및 어미와 같은 순수 문법 형태소를 제거하고 복합명사 분해 및 연속된 명사구 묶음 통해 검색 정확도를 높일 수 있었다. 색인 효율성 및 속도 개선을 위해 B+ tree 색인 알고리즘을 자체 개선하였으며, 다국어 CDML 및 질의어 분석을 위해 언어별 고급 분석 모듈이 플러그인 될 수 있도록 설계하였다. 색인 순서는 Fig. 4와 같다.

1) 띄어쓰기 비종속 형태소 분석기

복합 문서 중 PDF파일의 경우는 다른 문서 형식과는 달리 문단 단위의 자료 구조가 아닌 라인 단위의 자료 구조를 가지고 있다. 또한 원본의 사용자 띄어쓰기 오류와 함께 PDF 변환 시 시스템에 의해 추가적으로 띄어쓰기 오류가 발생하는 경우가 종종 발견되고 있다. PDF 문서의 색인 시 기존의 문장 절단 및 띄어쓰기 오류를 보정하지 않은 형태소 분석기를 사용할 경우 그 분석률이 현저하게 떨어지는 문제점이 발견되었으며, 이러한 문제를 해결하기 위해 새로운 통계 기반의 띄어쓰기 비종속 형태소 분석 및 품사 태깅 시스템을 자체 개발 및 적용하였다.

2) 색인 방식

일반적인 텍스트 검색과 달리 개체에 대한 정보도 같이 검색되어야 하기 때문에 색인 데이터가 여러 부문으로 분리되어 있다. 개체에 대한 검색은 속성 타입으로 분리하여 AND, OR, NOT 등의 다양한 검색이 가능하다.

3. 검색 결과 출력 시스템

검색 시스템은 크게 질의 분석기와 검색엔진 호출 및 검색 결과문 생성 시스템으로 구성되어 있다. 질의 분석기는 색인 시스템에서 사용했던 띄어쓰기 비종속 형태소 분석기를 사용하고 있으며, 형태소 분석기에서 추출된 중심 형태소를 검색 엔진에 질의함으로 검색 결과를 획득하게 된다. 검색 엔진에 질의 시, 개체의 종류, 날짜와 같은 각종 속성 조건을 부여할 수 있으며, 검색 키워드에 *, ? 등의 부분 검색 및 인용부호를 사용한 전문 검색(Full Text Retrieval), 괄호 등이 사용된 복잡한 질의문의 사용이 가능하다.

위에서 설명한 분석모듈과 색인 모듈, 검색 모듈은 서로 간의 영향을 주지 않도록 하기 위해서 서로 다른 프로세스 영역에서 동작하도록 하였으며 전체 프로세스들을 한꺼번에 모니터링할 수 있는 관리자 프로그램을 개발하여 제품의 상용성을 높였다. 이 관리자 프로그램은 예기치 못한 시스템의 오류 발생 시 자동으로 검색 시스템을 재시동하고 복구를 수행함으로 시스템의 안정성과 신뢰성을 높여주고 있다. 본 개체 검색 시스템의 특징 중 하나는, 기존의 검색 시스템이 검색 요청한 키워드가 특정 문서에 포함되어 있음을 단편적으로 알려주는 한계를 극복하여, 해당 문서의 어느 위치에, 어떤 문장에 몇 번이나 나왔는지 확인할 수 있도록 상세 검색 결과를 제공한다는 것이다. Fig. 5와 같이 키워드가 포함된 문단을 모두 추출하거나, 문서편집기 내에서 문서내의 해당 위치로 자동 이동하는 기능을 제공하고 있다. 상세 검색 결과인 문서내검색의 웹 화면과 사용자의 다양한 문서 편집기와의 연동을 위해 ActiveX 컨트롤이 사용되었다.

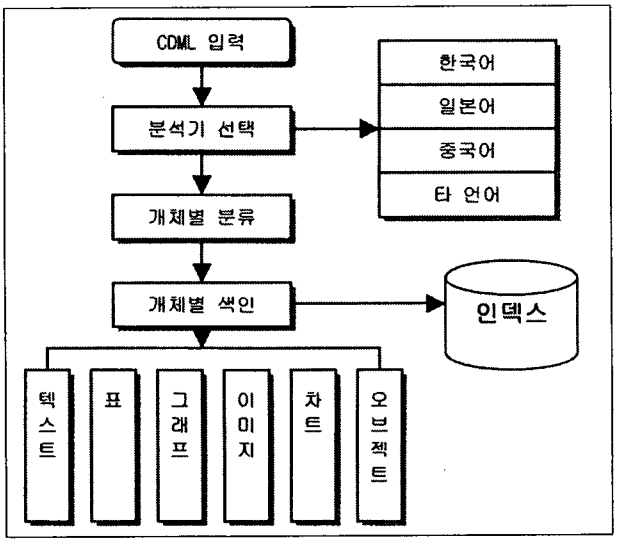

Fig. 4. 색인 시스템.

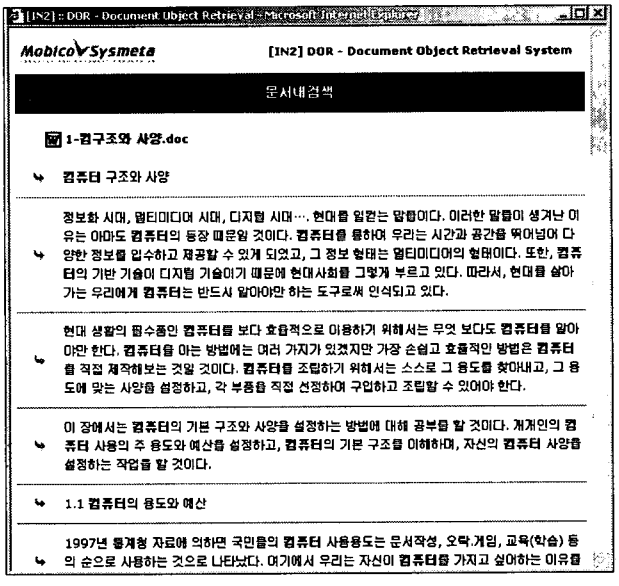

Fig. 5. 키워드 발생 구문 모두 보여 주기.

4. 문서 내 자동 위치 이동 기능

기존 검색 시스템의 경우, 검색된 문서에 존재하는 키워드 포함 문장을 찾아내기 위해 문서를 다운로드 받은 후에 다시 해당 문서 편집기 내부의 검색 기능을 사용해 별도로 키워드를 검색해야 하는 불편함이 있었다. 또한 특정 문서에 그래프나 그림과 같은 개체가 존재함을 알고 있더라도 이러한 개체를 찾기 위해서는 처음부터 일일이 문서를 검토해 봐야 하는 불편함이 있었다. 이러한 문제를 해결하기 위해 [IN2] DOR은 문서 내의 모든 개체 위치를 추출하고 색인 데이터에 함께 저장함으로, Fig. 6과 같이 검색 완료 후 실행된 해당 문서 편집기 내에서 찾고자 하는 키워드의 위치로 자동 이동할 수 있는 기능을 구현하였다. 이 기능을 구현하기 위해 ActiveX 컨트롤을 사용하였다. 검색 결과를 표시하는 웹 문서 내에서 스크립트 형태로 ActiveX 컨트롤에 키워드 위치를 전달하며, ActiveX 컨트롤은 검색된 문서의 편집기를 실행시킨 후, 이 편집기를 제어하여 찾고자 했던 키워드 혹은 문서 개체의 위치로 자동 이동하고 하일라이팅 표시를 한다. ActiveX 컨트롤은 각 편집기 별로 다양한 형태의 인터페이스를 통해 문서편집기와 연동한다. 예를 들어 오피스의 경우는 Automation 기술을 사용하여 제어하고, HWP 815 버전의 경우 DDE를 사용하고 있다. 이 동작 과정은 Fig. 7과 같다.

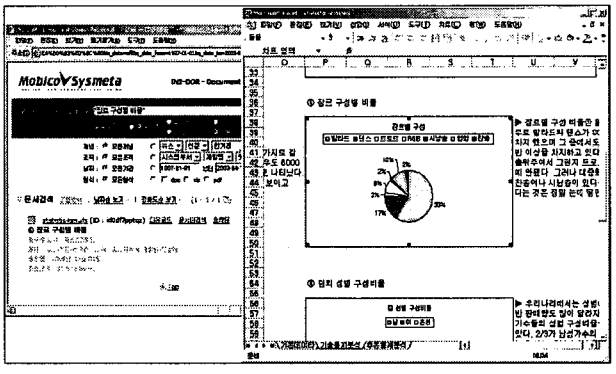

Fig. 6. 엑셀 문서에서 차트 개체 검색 및 위치 하이라이트.

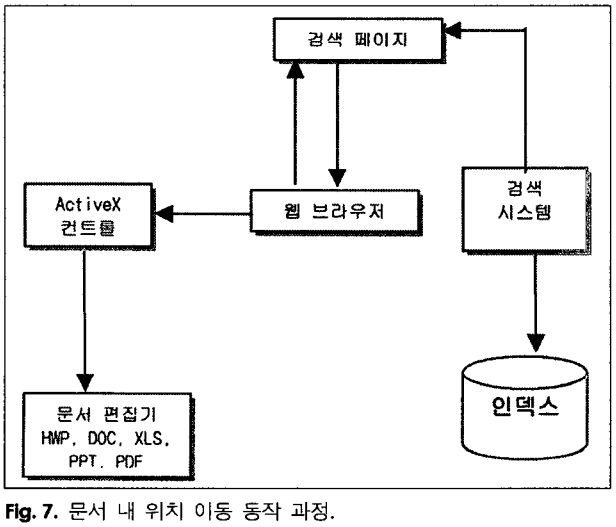

Fig. 7. 문서 내 위치 이동 동작 과정.

문서 개체 분석 및 색인 성능 평가

전체 시스템 프레임웍은 JAVA를 사용해 개발 되었으며, 문서에 대한 변환 모듈은 JAVA와 C++를 사용하여 제작되었다. 실험은 우선 총 25,242개의 복합문서를 인터넷에서 무작위로 수집하여, CDML 변환에 대한 성능 테스트를 수행하였다. 시스템은 P-III 633, 512MB, 60GB IDE HDD, Windows 2000 PC에서 수행하였으며, 속도는 OS 내부에 있는 시간관련 함수를 사용하여 측정하였다.

실험 결과는 Table 1과 같다. 변환 성공률은 대략 95~99%정도를 보이고 있으며, 인터넷에서 무작위로 수집한 문서이기 때문에 문서가 손상된 경우와 바이러스가 감염된 것 등이 분석 오류의 주 원인으로 분석 되었다.

결 론

본 논문에서는 가장 보편적으로 사용되고 있는 문서 포맷인 오피스, PDF, HWP 문서에 대하여 텍스트뿐만 아니라,

Table 1. 복합문서 별 XML변환 필터 성능 실험 결과

문서	doc	xls	ppt	pdf	hwp
전체문서	1491개(640MB)	595개(111MB)	2529개(5.8GB)	6181개(6.2GB)	14446개(3.2GB)
성공문서	1481개(638MB)	581개(107MB)	2384개(3.7GB)	5924개(4.8GB)	14346개(3.2GB)
XML문서	95.8MB	124.4MB	29.0MB	192.5MB	805.3MB
성공률	99.3%	97.6%	94.3%	95.8%	99.3%
소요시간	173초	210초	272초	9916초	8160초
복합문서/시간	3.7MB/초	8.6개/초	500KB/초	2.7개/초	13.6MB/초
XML문서/시간	553.8KB/초	591.8KB/초	107KB/초	19.5KB/초	98.7KB/초
비고	1. 프로세스 : 복합문서 -> 텍스트 -> XML 2. 변환 시 표, 그래프 등의 오브젝트 추출 포함 3. 실패 문서의 경우, 지원하지 않는 문서 포맷 및 깨진 문서가 대부분임				

표, 이미지, 그래프와 같은 각종 개체에 대한 검색이 가능하도록 복합 문서에서 CDML로 변환하는 방법과 해당 개체의 위치 정보를 이용한 문서 내 자동 하이라이팅 기능 구현 방법에 대하여 기술하였다.

기존의 텍스트에 기반한 문서 검색 방식의 한계를 넘어, 다양한 문서 개체에 대한 검색 방법을 제안함으로, 향후 검색 시스템의 새로운 패러다임이 될 수 있을 것으로 기대된다. 현재는 EDMS 및 KMS와 같은 다양한 문서/지식 관리 시스템의 기본 검색 시스템으로 활발히 적용되고 있으며 100만 건 이상의 방대한 다국어 문서에 대한 검색 성능 및 안정성 테스트를 완료한 상태이다. 앞으로는 자사의 텍스트 마이닝 시스템과 결합하여 보다 높은 수준의 정보 검색 및 분석 시스템으로 발전시켜 나갈 계획이다.

REFERENCES

류범종, 윤화묵(2002) : "인터넷 기반의 복합문서 자동화 시스템에 관한 연구", 데이터베이스연구

텀즈 페이지 : http://www.terms.co.kr

W3C 페이지 : http://www.w3c.org/

Croft WB : *"Information Retrieval System : Theory and Implementation", Kluwer Academic Publishers*

허용도 : "Java를 이용한 인터넷 정보 검색기 설계 및 구현", 산학기술협력연구 논문집 제 3권

백광진, 김태윤 : "자바에 기반한 웹 정보 검색 에이전트의 설계", 정보과학회 학술발표논문집 제 25권 1호

"Microsoft Office Development with Visual Studio", Microsoft Corporation

한경수, 이도길, 임해창 : "통합정보 검색을 위한 과학 기술 문서 색인 및 요약 시스템의 개발"

Adobe Acrobat 도움말 파일, 2003