한국컴퓨터정보학회 동계학술대회 논문집 제19권 제1호 (2011. 1)

# XML 기반의 본문검색 미들웨어 시스템 설계 및 구현

김효남°
°청강문화산업대학 e스포츠게임과
e-mail: hnkim@ck.ac.kr

# Design and Implementation of the Search Inside Middleware System by using XML

Hyo-Nam Kim°
°Dept. of e-Sports Game, ChungKang College of Culture Industries

● 요 약 ●

최근 스마트 디바이스 기반의 다양한 콘텐츠 제작 공급에 대한 새로운 마켓 시장 형성과 태블릿 디바이스 기반의 전자책 시장 규모의 확대에 따른 새로운 유형의 디지털 콘텐츠 시장이 형성되고 있다. 디지털 미디어는 정보환경 범위를 종이의 범위에서 벗어나 매체와 연계한 다양한 형태로의 변화뿐만 아니라 서술 방식과 소통방식의 방법에도 변화를 발생하였다. 그리고 지면에서 국한된 아날로그 매체의 물리적, 공간적, 시간적인 한계를 뛰어넘어 다양한 콘텐츠를 손쉽게 접근할 수 있게 만들었다. 이런 환경에서 본문검색 서비스는 아날로그와 디지털 매체의 상호 공존관계를 형성할 수 있다. 본 논문에서는 그림종이문서를 본문검색이 가능한 이미지형태의 디지털문서로 변환해주는 디지털라이징 시스템으로 문자위치정보를 포함하는 광학문자인식(OCR)기능과 인식된 문자의 오류를 수정하는 에디터기능을 통해 추출된 내용을 XML형태로 제공하는 본문검색 시스템을 제안한다. 특히, 문자인식 후처리 공정에서 복수의 관학문자인식(OCR)엔진을 통해 결과 비교와 문자위치 정보 확인 및 편집, 맞춤법 검사 등의 특화된 기능 등은 본  논문에서 가지는 강점으로 디지털문서 구축에 소요되는 시간과 비용을 혁신적으로 절감시켜준다.

**키워드:** 본문검색(Search Inside), 광학문자인식(Optical Character Reader), XML

## I. 서 론

현재 인터넷 시장은 브로드밴드 인프라의 확산과 새로운 스마트 디바이스 기반의 다양한 콘텐츠 제작 공급에 대한 새로운 마켓 시장 형성 그리고 융복합 기술 기반의 비즈니스 모델 전환이 빠르게 이루어지고 있는 것이 오늘의 상황이다. 이런 시장 상황에서 온라인포털 시장의 흐름도 매우 빠른 속도로 변화하고 있으며, 온라인포털 시장이 검색 광고 분야에서 큰 성과를 보이면서 온라인포털의 시장 규모 및 전망이 밝다고 본다. 온라인포털 시장이 밝은 이유 중에 하나가 지식의 대중화를 목적으로 하는 포털의 기본 취지를 가능하게 하는 것이 책에서 비롯되는 출판콘텐츠의 디지털화이다[1].  디지털 미디어의 특성은 정보환경 범위를 종이의 범위에서 벗어나 매체적인 변화뿐만이 아니라 서술 방식과 소통방식의 방법에도 변화가 발생하였으며, 지면에서 국한된 아날로그 매체의 물리적, 공간적, 시간적인 한계를 뛰어넘어 다양한 정보에 손쉽게 접근할 수 있게 만들었다. 이런 환경에서 출현한 도서본문검색 서비스는 아날로그와 디지털 매체의 상호 공존관계를 형성하고 있다[2].

우리나라에서 본문검색(Search Inside)은 도서의 경우 2004년에 시작되었고, 잡지의 경우 2007년에 시작되었다[3]. 본문검색은 포털사이트에게는 검색의 신뢰성을 높여주는 중요한 방편으로 인식되었고 출판사와 잡지사에게는 책과 잡지의 홍보효과를 높여줄 수 있을 것이라고 생각하고 있다.

국내 행정DB구축사업과 지식정보화 사업 등에서 디지털화하는 부분에 역점을 두고 있으나 산출물인 이미지에 본문검색 기능을 활용한다면 대국민 서비스가 원활이 이루어지며 DB구축사업의 한 단계 질을 올릴 수 있을 것이다. 매일 출간하는 도서간 200종인데 도서를 제작할 때 사용한 파일은 실시간 수집이 불가능 하여 일반 도서를 스캔하여 이미지 상태에서 텍스트 축출을 하여 효율적인 본문 검색 기술을 적용한다면 실시간 도서 지식 검색이 가능하여 포털 사이트에서 필수적인 미들웨어 시스템이 될 수 있습니다.

본 논문에서는 기술적으로 적용할 수 있는 범위는 아마존, 구글, 네이버 같은 지식 포털 사이트에서 본문 검색을 위한 필 수 프로그램으로 특히 한국어처럼 2바이트를 사용하는 국가에서는 절대적으로 필요한 프로그램이 될 수 있다. 기술 내용은 스캔된 이미지 데이터를 다중 OCR처리 후 동일한 데이터만 축출하고 동일하지 못한 데이터를 XML편집기를 이용하여 수정 편집 할 수 있는 미들웨어 시스템을 소개한다.

## II. 관련 연구

### 1. 관련연구

#### 1.1 국내 동향

국내 인터넷의 도서본문검색 서비스는 2004년 전자책 전문업체 북토피아, 네이버, 2006년에는 다음, 엠파스, 파란, 싸이월드 등의 포털에서 실시하고 있으며 교보 온라인서점, 예스24, 인터파크 등에서도 실시하고 있다[4]. 특히 본 연구와 기술적으로 밀접하게 관련 있는 본문검색 소프트웨어 솔루션으로 퍼셉컴(주)사의 아르미 6.0과 한국인식정보기술의 Hi_글눈에 대한 특징을 살펴보면 다음과 같다. 아르미6.0은 순수 Text 스캔 파일로 광학문자인식(OCR: Optical Character Reader) 84%의 정확도를 가지고 있으며, 이미지와 Text 복합 스캔 파일은 광학문자인식 60%의 정확도를 갖는다. 그리고 텍스트/이미지 레이어 영역 분리 기능을 가지고 있지 않다. 다음은 Hi-글눈으로 순수 Text 스캔 파일의 광학문자인식 80%의 정확도를 가지고 있으며, 이미지와 Text 복합 스캔 파일인 경우 광학문자인식 75%의 정확도를 갖는다. 그리고 텍스트/이미지 레이어 영역 분리 기능을 가지고 있지 않다.

#### 1.2 국외 동향

출판콘텐츠의 대표적인 전자책(eBook)의 세계적인 시장규모는 아래 그림1에서와 같이 2007년 Amazon Kindle이 부상하면서 전자책(e-Book) 시장성장의 견인역할을 수행하며, 세계 전자출판 산업은 2014년까지 연평균 27.2%씩 성장, $82억 6,200만 규모의 시장을 형성할 것으로 전망되고 있다[5]. 전자책을 기반으로 도서본문검색 기능이 강화된다면 시장의 규모는 더 늘어날 것으로 예상한다.

(단위: $백만)

| 구 분 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | '09~'14 CAGR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 북미권 | 768 | 907 | 1,077 | 1,301 | 1,470 | 1,804 | 2,219 | 2,701 | 3,260 | 3,890 | 21.5% |
| 유럽권 | 15 | 30 | 68 | 17 | 146 | 197 | 292 | 519 | 837 | 1,257 | 53.8% |
| 일본 | 70 | 168 | 243 | 327 | 358 | 397 | 454 | 529 | 626 | 746 | 15.8% |
| 중국 | 3 | 11 | 21 | 41 | 51 | 67 | 102 | 157 | 224 | 304 | 42.9% |
| 아태권 | 27 | 82 | 170 | 311 | 349 | 424 | 561 | 885 | 1,225 | 1,585 | 35.3% |
| 남미권 | 9 | 26 | 52 | 94 | 103 | 123 | 162 | 257 | 360 | 480 | 36.0% |
| 합 계 | 892 | 1,224 | 1,631 | 2,191 | 2,477 | 3,006 | 3,790 | 5,048 | 6,532 | 8,262 | 27.2% |

자료: 한국콘텐츠진흥원(2010)

그림 1. 세계 전자책 시장의 규모

Fig. 1. Global Market of the e-Book

국외 인터넷 도서본문검색 서비스의 대표적인 검색 포털 기업은 구글이다. 구글은 2005년 11월 출판업계의 반발에도 불구하고 '구글 프린터(Google Print)'라는 도서검색 프로그램을 소개하였다[6]. 현재 구글의 검색엔진에서 원하는 키워드를 입력하면 수 천권의 책이 스캐닝한 원본 이미지를 볼 수 있다. 구글 프린터는 본문검색의 대표적인 기술이다. 그 이외로 국외 본문검색 솔루션으로 IRIS사의 Readdiris와 ABBYY사의 Fine Reader8.0을 특징을 알아보면 다음과 같다. Readiris는 순수 Text 스캔 파일인 경우 광학문자인식 90%의 정확도를 가지며, 이미지와 Text 복합 스캔 파일은 광학문자인식 80%의 정확도를 보이고 있다. 텍스트/이미지 레이어 영역 분리 기능도 가지고 있으며, 또한 아시아 권 2바이트도 지원한다. Fine Reader8.0은 순수 Text 스캔 파일인 경우 광학문자인식 90%의 정확도를 가지며, 이미지와 Text 복합 스캔 파일은 광학문자인식 80%의 정확도를 보이고 있다. 텍스트/이미지 레이어 영역 분리 기능도 가지고 있다.

## III. 본 론

### 1. 본문검색 시스템 설계

아래 그림 2는 본 연구에서 제안하는 본문검색을 위한 시스템 구성을 보여주고 있다. 본 연구에서 소개하는 본문검색 시스템은 e-Book Encoder, e-Book Character, e-Book DRM(Digital Rights Management), Search Engine 등 4개의 주요 시스템으로 구성되어 있다.

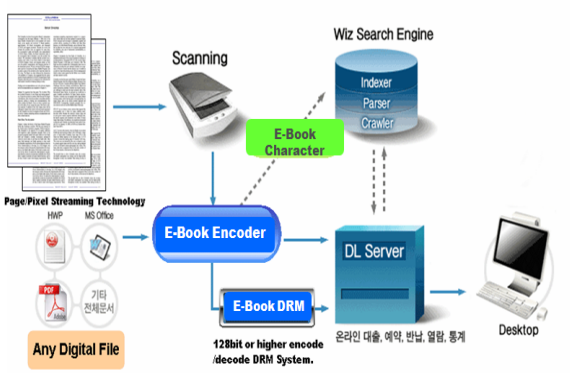


그림 2. 본문검색 시스템 구조

Fig. 2. Search Inside System Structure

e-Book Encoder는 원문을 스캐닝한 데이터를 encoding하여 본문검색의 기본 데이터를 구성하게 하는 시스템이다. e-Book Character는 본 연구에서 제안하고자 하는 주요 내용으로 본문검색이 가능한 이미지형태의 디지털문서로 변환해주는 디지털라이징 시스템으로 문자위치정보를 포함하는 문자인식(OCR) 기능과 인식된 문자의 오류를 수정하는 에디터기능을 통해 추출된 내용을 XML형태로 제공하기 위한 시스템이다. e-Book DRM은 콘텐츠가 정해진 규칙 내에서만 사용되는 기술로서 특정 사람이나 하드웨어에서 보기만 가능, 보기/인쇄만 가능, 정해진 횟수 및 기간만큼만 사용, 열람 내역 추적과 같은 본문검색의 제어기능을 제공하는 시스템이다. Search Engine는 OS 에 종속적이지 않은 시스템으로 Unix, Linux, Windows Server를 지원하며, Windows 계열 서버에는 엔진에 접근 가능한 인터페이스를 제공한다. 색인 크기는 원본 문서의 50% 정도(본문 전체 검색일 경우)이며, 하나의 색

인 데이터에서 DB , File , Web 등의 모든 문서에 대한 검색을 한 번에 할 수 있다. DB Crawler , File Crawler Robot 들이 수집한 원본을 Indexer 가 이해 할 수 있도록 XML 형태로 원본 문서를 변환 하여 Indexer에 제공할 수 있게 하는 XML 방식의 데이터 공유를 지원한다. 키워드 검색의 장점과 각종 검색 연산자 (+,-, AND , OR 등)를 지원 엔진 내부에서 본문의 데이터를 저장하여 따로 데이터를 저장 할 필요가 없도록 하였다.

## 2. e-Book Character 시스템 구현

그림 3은 본문검색을 위해 원문 자료에서 본문검색용 XML 자료의 변환까지의 공정을 보여주고 있다.


그림 3. 본문검색용 XML 데이터 생성과정
Fig. 3. Process of XML Data Generation for Search Inside

첫 번째 공정으로 본문검색에 사용될 원문을 스캐닝하기 위하여 원문 자료를 해체하여 스캐닝하기 위한 준비를 갖는다. 스캐닝 및 이미지 공정 과정은 해체된 원문을 스캐닝과정을 통해 생성된 이미지를 본문 검색을 위한 교정 작업을 수행한다. 이미지 교정 작업은 크게 3가지 작업을 수행한다. 첫 번째로 센터링 작업으로 이미지의 계단 현상을 없애기 위한 센터링, 센터링 하고자하는 기울기를 사전 조정, 특정 폴더 내의 모든 문서를 일괄 센터링하는 작업이 포함된다. 두 번째로 영역 센터링으로 문서의 일부 영역만을 선택해서 그 부분이 A4의 가운데에 위치하도록 하고 특정 영역을 잘라낸 후 센터링 한 효과가 나도록 한다. 세 번째 작업은 이미지 이동으로 문서가 한쪽으로 치우친 경우 이미지를 이동시키는 작업이다.

문자인식 전처리 공정에서는 종이문서를 스캔  이미지로부터 문자내용과 위치정보를 자동적으로 식별 추출하는 공정으로 복수의 OCR엔진을 적용하여 인식결과의 정확성을 향상시킨다.

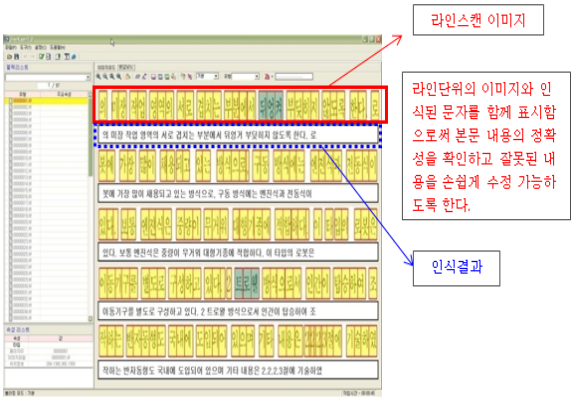


그림 4. OCR 후처리 공정
Fig. 4. OCR Post-Processing

그림 4의 문자인식 후처리 공정에서는 추출된 문자내용과 위치정보를 편집하여 가공하는 공정으로 직관적인 문자 위치 정보의 확인 및 편집 기능 과 복수의 OCR엔진 비교로 한번 걸러진 본문 내용을 맞춤법 검사 기능을 통하여  수정함으로써 최종 산출물에 대한 Quality를 확보한다.

그림5의 XML 생성공정은 정의된 문법(Schema, DTD)에 따라 XML 파일을 생성하는 것으로 최초 문자인식 공정에서 자동으로 생성 되며, 이후 공정에서는 직접적인 XML의 편집 없이 유저 인터페이스에 의해 수정되므로 XML문서의 유효성(Validation)을 손쉽게 관리할 수 있다.

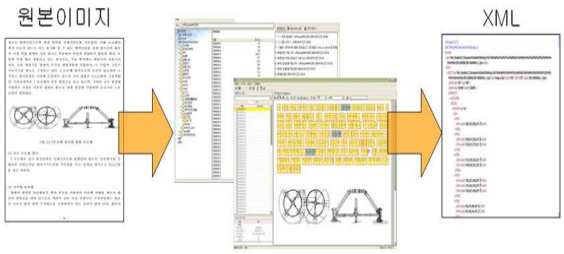


그림 5. XML 데이터 생성
Fig. 5. Formation of the XML Data

본 논문에서 제안하는 본문검색 시스템은 아날로그 자료(종이/문서/도서/고도서)를 디지털화한 후 산출물인 이미지 파일을 이용하여 90%이상의 Text를 축출하여 교정 수정되는 Text의 위치 정보를 XML으로 구현하고 위치 좌표를 수정 변경 할 수 있는 통합 XML 편집기 구현했으며, 스캔된 이미지 파일을 OCR적용 중 이미지와 텍스트 구역을 자동 인식하여 텍스트 영역만 OCR이 구현되도록 하고 이미지를 OCR하여 생성되는 일명 '쓰레기 데이터'가 생성되지 않게 하는 OCR자동 구현 기능을 구현하였다.

## Ⅳ. 결 론

새로운 스마트 디바이스와 태블릿 PC 기반의 다양한 콘텐츠 제작 공급에 대한 새로운 마켓 시장 형성과 전자책 시장 규모 확대에 따른 새로운 유형의 디지털 콘텐츠 시장이 형성되고 있다. 본 연구에서는 전자책에 응용할 수 있는 효율적인 본문검색 시스템으로 그림종이 문서를 본문검색이 가능한 이미지형태의 디지털문서로 변환해주는 디지털라이징 시스템이며, 문자위치 정보를 포함하는 광학문자인식(OCR) 기능과 인식된 문자의 오류를 수정하는 에디터기능을 통해 추출된 내용을 XML형태로 제공하여 디지털 문서 구축에 소요되는 시간과 비용을 절감시켜줄 수 있는 시스템이다. 향후 새로운 디바이스 기반의 전자책 시스템으로 적용할 수 있는 효과적인 본문검색 시스템 연구가 필요하다.

## 참고문헌

[1] Young, N. D. “How Digital Content Resellers are Impacting Trade Book Publishing.” Springer. Pub Res Q(2009) 25:139~46. 10 June 2009.

[2] Clarida, R. W. “Electronic copyright Rights: Do You Have What You Need?” Springer. Pub Res Q(2009) 25:199~204. 3 November 2009.

[3] Yong Jun Lee “A Study on the Search Inside of books and magazines” Proceeding of the 2008 Korean Publishing Science Society Conference pp.221~251, 2008.

[4] Lee, Chan Hee(2007) “Study on the Usefulness of the Online Book Text Retrieval Servic” a thesis for a Master's degree, Chung-Ang University, pp.26~27, 2007

[5] Treanor T. “Amazon: Love Them? Hate Them? Let’s Follow the Money.” Springer. Pub Res Q(2010) 26:119~128. 2 June 2010.

[6] MYDAILY News Paper, http://www.mydaily.co.kr/24 March 2006.