전문에 대한 검색시스템의 구현

김 대규′ , 정 희택″ , 강 영만″ , 한 순희″ , 조 혁현″
′ 순천대학교, ″ 여수대학교

Implementation of Information Retrieval System for Full-Text

Dae-Ku Kim′ , Hee-Taek Ceong″ , Yong-Man Jang″ , Soon-Hee Han″ , Hyug-Hyun Cho″
′ Sunchon National National University, ″ Yosu National University
E-mail : dkkim@cs.snu.ac.kr,{htceong,ymjang,shhan,hhcho}@cs.yosu.ac.kr

요 약

인터넷을 이용한 정보검색이 일반화되면서, 보다 정확하고 꼭 필요한 정보의 요구가 일반화되었다. 정확한 정보의 제공을 위해, 요약된 정보에 대한 중심어(keyword) 검색뿐만 아니라 전문(Full-Text)에 대한 검색 요구가 일반화되었다. 본 연구에서는 전문 검색을 위한 설계 방안을 제안한다. 기존에 제안된 전문 검색 방안과 오라클에서 제공하는 interMedia Text를 이용한 전문 검색 방안을 비교한다. 이를 기반으로 정보 검색 시스템에서 구현 방안을 제시한다.

ABSTRACT

Using the Information Retrieval systems on the Internet, the demand of exact and specific information has also been popularized. To offer exact information, there has been generalized demand of searching from the keyword of the shortened text and also of the full-text. This study is to suggest a scheme for full-text searches. It is to compare the existing scheme of information search and full-text information search with interMedia text. We suggest search methods for the full-text.

Ⅰ. 서 론

다양한 문서편집기의 보편화에 따라, 국내 대학 및 여러 기관, 연구소에서는 전문 데이터 베이스를 구축하고 있다. 전문은 이전의 고전적인 텍스트 형태의 정형적인 문서정보로부터 비디오, 오디오, 이미지 등의 다양한 멀티미디어 정보들로 구성하고 있다. 더욱이 인터넷 환경의 급속한 확산과 사용자들의 전문 데이터 서비스에 대한 요구의 증가는 계속되고 있다. 이러한 요구를 만족하기 위해서는 여러 형태의 복합문서에 대한 변환 및 처리방법에 연구되어야 한다[1]. 기존 시스템들은 고전적인 텍스트 형태의 전문 데이터 베이스를 구축을 기반으로 하며 단순한 키워드를 기반으로 한 인터넷 서비스를 지원하고 있다. 그러나 다양한 문서형식이나 스타일에 대한 표준화 부재와 멀티미디어 형태의 정보를 수용할 수 없음으로 인해, 인터넷을 통해 효율적인 서비스를 제공되고 있지 못하다.

본 연구에서는 현재 활용되지 못하고 있는 멀티미디어 정보의 전문과 방대한 양의 전자형태의 논문을 오라클에서 제공하는 interMedia Text를 이용한 하나의 전문 데이터 베이스를 구축하고 이를 활용하여 기존에 제안된 전문 검색 방안과 본 연구에서 제안 하고자 하는 전문 검색 방안을 비교한다. 이를 기반으로 정보 검색 시스템에서 구현 방안을 제시한다.

본 논문의 구성은 제 2장에서 현재 인터넷의 현황과 차세대 인터넷 액세스 환경을 설명하고 전문 검색 시스템의 필요성을 제시한다. 제3장에서 기존의 전문데이터 베이스의 구성과 전문검색 방안을 설명하고 문제점을 파악한다. 제4장에서는 앞서 제시한 문제점을 interMedia Text를 이용한 전문 검색 시스템 구현 방안을 제시하고자 한다.

Ⅱ. 본 론

1. 현재 인터넷 과 차세대 인터넷 액세스

과거 인터넷은 전문가들에게만 인식되고 사용되어 왔지만 현재 인터넷은 일반 사람들에게 매우 친숙한 통신매체로 자리잡아가고 있다. 국내 인터넷 액세스망은 크게 ADSL(Asymme- tric Digital Subscirber Line), HomePAN(Home Phoneline Network Alliance), 케이블 모뎀 등으로 구분되어지고 급속도로 이런 초고속 인터넷의 가입자가 증가하고 있다[2]. 하지만 이 또한 인터넷의 급속한 발전과 광대역 서비스 요구로 광대역 인터넷 기술을 가져왔고 그에따라 빠른 미래에는 VoDSL(Voice over DSL), VDSL(Very high bit rate DSL), 디지털 STB(Set Top Box) 등 이러한 차세대 인터넷 액세스 기술이 선보이고 있다[3].

2. 인터넷과 전문 검색 시스템

최근까지 온라인 상용시스템에서 전문 데이터베이스가 끊임없이 증가하여 왔고, 최근에 온라인 데이터베이스 중에서 전문 데이터 베이스가 차지하는 비율이 크게 증가하고 있다.

인터넷환경에서의 자동전문검색시스템은 많은 이점을 제공한다. 첫째, 정보 및 컴퓨터 기술의 끊임없는 발전은 보다 빠르고 저렴하며, 신뢰도가 높고 사용하기 편리한 컴퓨터로의 접근을 가능하게 하고 있다. 둘째, 자동화된 색인 기법에 의해 일관성 있는 정보의 유지 및 접근을 가능하게 한다. 셋째, 정보원인 전문데이터는 출판을 위해 원고를 준비하는 과정에서 쉽게 획득할 수 있다. 마지막으로, 문헌의 전문을 탐색함으로써 보다 정확하게 원하는 결과에 접근할 수 있다[4].

인터넷을 통해 많은 사용자들이 간단하고 요약된 서지 정보보다는 온라인으로 원문을 제공받기를 원하고 있지만 현재 제공되는 서비스에는 한계가 있다. 물론 본 연구를 위해 참고문헌 중에 전문 검색 시스템의 구축 방안을 제시하고 있지만, 현재 국내에서는 아직까지 전문 검색 시스템의 이용 및 효율성연구는 초기단계에 있으며 끊임없는 연구가 이어져야 할 것이다.

III. 기존의 전문 검색 시스템

인터넷을 통해 특정한 정보를 제공하고자 노력은 통신 분야의 발전과 더불어 정보를 가공 처리하여 서비스하는 부분에 많은 노력을 경주하고 있다. 정보의 가공 처리 대상으로써, 다양한 형태의 논문을 포함하며 이를 전문 데이터 베이스화 하고자 한다. 대상으로써 전자 문서의 경우는 규준형태(canonical form)를 결정하는 부분에서부터 데이터 베이스에 정보를 저장하는 방법, 검색을 위해 색인 하는 방법이 다양하다. 현재 미국의 Adobe사의 PDF 파일 형식이 인터넷을 통해 가장 많이 전자 문서화되어 서비스되고 있다[1]. 그 외에도 <표 1>과 같이 일반적으로 많이 사용되는 규준형태를 보여주고 각 형식에 대한 개요 설명, 적합성, 본문 검색, 원본과의 동일성대한 비교를 보여준다[5].

<표 1> 검색시 통계치 비교

형식	개요	적합성	본문검색	원본과의 동일성
DVI	Device Independent 형식	적합	가능	비교적 동일, 서체대체
PS	PS 프린터와 그래픽을 위한 형식	영문은 적합	가능	비교적 동일, 서체대체
PDF	Adobe가 만든 압축 PS형식	부적합	가능	동일
DOC	MS Word 문서	부적합	개발가능	동일
HWP	아래아 한글 문서	부적합	개발가능	동일
XLX	레이저젯 프린터를 위한 PCL에기반	제한적으로 적합	HWP는 이미지	흑백, A4
XML	SGML의 실현 가능한 형식	당장은 부적합	가능	?

일반적으로 각 대학에서 그 대학의 석박사 학위논문들이 아니면 전자형태의 자료를 보유하고 있지 않다. 그래서 이미지형태의 자료만을 서비스하고 있는 부분이 일반적이다. 이러한 형태의 전문 데이터가 현재 보편적인 전문 데이터 베이스 형태이다.

이렇게 구성된 데이터 베이스를 일반적인 검색 시스템 통해 서비스하기 위해서는 저자명, 제목 등 몇몇 중요 필드를 인덱스 처리하거나 필드 내 에서 검색하여 정보를 찾아내야 한다. 이렇게 구성되어진 시스템은 사용자들의 검색 과정에서 속도가 현저하게 떨어진다. 인덱스 처리는 많은 정보의 조회를 좀 더 빠르게 수행하기 위한 중요한 작업이다. 그러나 특정 몇몇 필드의 정보를 인덱싱 하였다고 그 정보에 대한 상세한 검색이 이루어졌다고는 볼 수 없다. 검색 시스템의 효율성을 높이기 위한 연구는 여러 분야에서 계속 진행중이며 높은 재현율을 유지하면서 정확성을 증진시키는 방법이 연구되어졌다[6,7]. 하지만 전문 전체를 인덱싱 처리를 하지 않는 이상 보다 정확한 정보를 검색하기는 어렵다.

위에서 언급한 일반적인 전문 데이터 베이스에서 검색 시스템의 이용은 사용자들에게 한정적인 서비스만을 제공할 수밖에 없다[8,9]. 또는 재현율만 높고 정확성이 떨어지는 정보만을 제공한다. 더욱이 멀티미디어 형태의 전문을 구축할 수 없고 이를 검색할 수 없다. 인터넷상에 존재하는 다양하고 많은 정보로부터 사용자가 원하는 정도가 가장 높은 정보의 제공은 필수요소다.

IV. interMedia Text의 전문 검색 시스템

일반적인 형태의 전문 데이터 베이스와 검색 시스템의 효율성을 높이고 정보의 조회를 좀 더 빠르게 수행하기 위한 방안을 제시하고자 한다. 그러기 위해 본 논문에서는 오라클에서 지원하는 interMedia Text를 이용한 전문 데이터 베이스의 구성과 검색 시스템을 제안 하고자 한다. <표 2>는 오라클 데이터 베이스 시스템을 이용해 일반적 구성 테이블과 interMedia Text를 이용한 테이블을 구성하고 한 필드의 크기를 600Byte로 고정하고 12750건의 데이터를 인덱스 하였을 경우 검색시 발생하는 통계치를 비교한 것이다.

<표 2> 검색시 통계치 비교

	일반적인Table		interMedia Table	
call	disk	query	disk	query
parse	0	0	0	0
Execute	0	0	0	0
petch	10620	12752	132	1492
total	10620	12752	132	1492

국내에서 구축된 데이터 베이스 시스템의 대부분은 오라클로 구축되어 있다. 이러한 이유에 의한 확장성과 데이터의 공유 측면에서 interMedia Text를 이용하였다. 본 연구에서는 기존의 시스템을 활용하여 보다 새롭게 가공처리 하지 않고 기 구축되어진 전문 데이터 베이스를 좀더 폭넓은 정보를 제공할 수 있는 방안을 제공한다.

인터넷 기반의 애플리케이션은 전자상거래 카탈로그, 기업 저장소(repository) 및 기타 풍부한 매체의 애플리케이션에서 사용되는 풍부한 데이타 타입을 지원하는 앞선 데이타 관리 서비스를 목적으로 개발되었다. 이미지, 오디오, 비디오, 로케이션, 텍스트 및 관계 데이터에 대한 액세스가 필요한 애플리케이션으로 확장된다.

interMedia는 다섯 가지 통합 컨텐트 관리 모듈을 갖고 있다. 첫째, ConText 언어 분석기술을 기초로 사용자로 하여금 우수한 검색 및 자연어 조회를 이용하여 쉽게 텍스트를 찾을 수 있게 한다. 둘째, 이미지를 관리하고 스케일링과 크로핑(cropping)과 같은 기본 조작 기능을 지원한다. 셋째 Oracle8i, Oracle Video Server, 웹사이트 및 특수 서버를 포함하는 다양한 소스로부터 브라우저 기반으로 영상 및 음향을 전달해준다. 넷째, 로케이터(locator)는 상점, 유통 포인트, 어떠한 주소로부터의 거리나 위치와 같은 공간적 변수를 기반으로 한 행사 등에 대한 정보의 위치를 알려준다. interMedia의 기능들은 RealNetworks의 RealVideo와 RealAudio 서버의 문서 관리 시스템, 지리적 정보 시스템 및 매핑 소프트웨어와 같은 소프트웨어들과의 통합 방안을 제공한다.. 이런 기능중에 ConText 언어 분석기술을 기초로 한 interMedia Text를 활용하여 인덱싱 처리를 수행한다. 다음 그림 1, 2와 같이 전문 검색 시스템에서 전문을 검색하여 재현율과 정확율을 높인다.

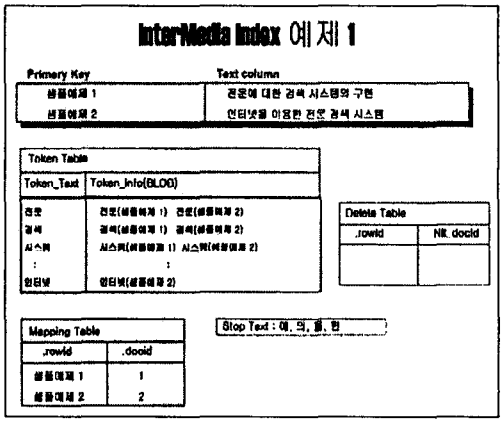

그림 1.인덱스예제1(생성)

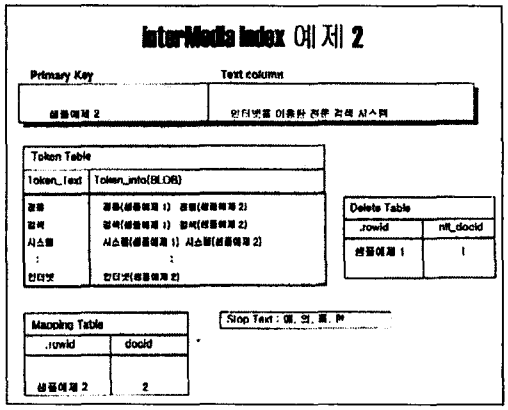

그림 2. 인덱스 예제2(삭제)

interMedia Text의 큰 특징을 살펴보면 하나의 SQL 질의로 데이타베이스 컬럼 내의 정보 뿐 아니라 문서의 본문과 주제를 검색 가능하게 한다. 또한, 주제별로 문서요약물을 생성할 뿐 아니라 문서 내에서 가장 강조되고 있는 주제를 식별한다. ASCII 텍스트, 워드, 엑셀, 파워포인트, WordPerfect, HTML, Acrobat/PDF 등과 같은 모든 일반적인 포맷에 저장된 문서들을 색인화 할 수 있다. 문서들을 HTML 포맷으로 변환하여 브라우저에서 간단히 볼 수 있고, 검색 단어 및 주제를 강조 표현한다. 모든 본문 검색과 분석 기능은 영어로 이용 가능할 뿐 아니라, 많은 다른 언

어를 지원하고, 본문 인덱싱, 본문 질의 및 기본 문자 지원 등과 같은 특정 언어에 요구되는 사항들을 지원하는 추가적인 특성들을 포함한다. 이런 기능들을 이용 전문 검색 시스템 구축한다면 오디오, 비디오, 이미지 형태의 데이터 타입과 연계하여 다양한 컨텐츠를 구축할 수 있다.

interMedia Text를 이용한 검색 시스템의 구성은 먼저, 그림 3과 같이 전자 형태의 논문 및 책자 형태의 논문을 가공하여 필터링 한다. 다음으로, interMedia Text Sever로 구분되어지는 논리적인 내부 구성환경에 따라 Text Dictionary와 사용자 정의 Table로 구성되고 전문 검색에 필요한 색인 작업이 이루어진다. 사용자가 인터넷을 통해 전문 검색을 지원하는 시스템에 접속하면 웹 미들웨어인 OAS(Oracle Application Server)에 의해 검색 엔진에 접근할 수 있다. 사용자는 찾고자 하는 논문의 특정한 필드의 선택에 따라 질의어를 입력한다. PL/SQL으로 작성된 검색 프로그램은 검색 엔진을 통해 interMedia 색인을 검색한다. 다음으로, interMedia Text Sever로 구분되어지는 논리적인 내부 구성환경에 따라 검색 질의어를 일반적인 검색 필드는 물론 전문까지 검색을 수행한다.

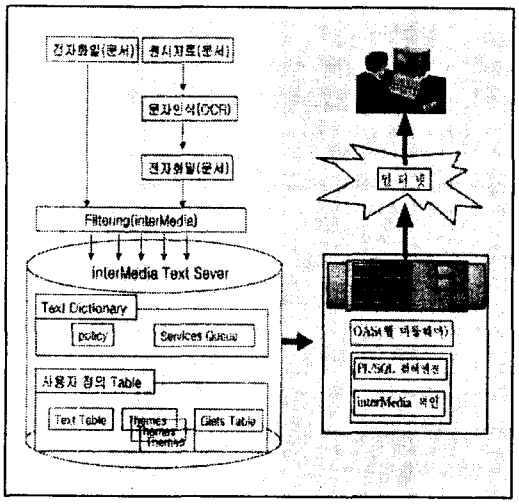

그림 3. 전문 검색 시스템 구성도

V. 결 론

본 논문에서 현재 각 대학 및 연구소에서는 전문 데이터 베이스를 구축하여 검색 시스템을 이용 인터넷을 통해 석박사 학위논문 및 연구 논문을 서비스하려고 한다. 이러한 추세는 계속 증가할 것이다. 하지만 전문 데이터 베이스의 활용을 좀 더 폭 넓게 이용할 수 있는 방안이 제시되어 있지 않다. 단순한 서지 정보 형태의 서비스나 데이터베이스에서의 키워드 검색 방안만을 제공하고 있다. 더욱이 다양한 데이터의 형태를 고려하고 있지 않다. 더욱이 사용자들만이 이용할 수 있는 형태의 검색 시스템에서 벗어나 멀티미디어 형태의 서비스를 구축하여 다양한 컨텐츠를 관리 운영할 수 있어야 한다.

본 논문에서는 전문 검색 시스템의 표준을 정하고자 한 것은 아니다. 다만 많은 기관에서 전문 데이터 베이스 구축과 검색 시스템을 구축함에 있어서 보다 효율적인 방법을 제안 하고자 한다. 다양한 컨텐츠와 다양한 형태를 대상으로 한 전문 검색 방안을 제공하고자 한다. 본 논문에서는 데이터 베이스 시스템으로 오라클을 기반으로한 interMedia Text를 이용하여, 기존의 전문 검색 시스템들의 문제점을 해결하고자 하였다. 다양한 언어를 포함하고 다양한 데이터 형태들을 저장 및 검색 대상으로서 제공하고자 한다. 이로인해 사용자에게 높은 재현율과 정확성을 보장하고자 한다. 더욱이 사용자에게 다양한 검색 서비스를 제공할 수 있다. 향후 연구에서는 전문 검색 서비스와 지리정보 시스템을 통합한 검색 서비스를 개발하고자 한다.

참고문헌

[1] 이기호, 김진숙, 윤화목 "학위논문 전문데이터 베이스 구축 및 서비스환경 구현", 제7권, 제1호, pp.41-49, 한국정보처리학회 논문지, 2000.1

[2] H. Schulzrinne, "World Wide Web: Whence, Whither, What next?", IEEE Network Mag., Mar./Apr. 1996

[3] 한운영, "차세대 인터넷 기술", 텔레콤, 제16권, 제1호, pp.52-57, 2000.

[4] 백인천, "CORBA 기반의 컴포넌트 기술과 전자 상거래 응용",제17권, 제7호, pp.29-36, 정보과학회지 1997

[5] 인터넷 전자문서의 형식 비교, http://www.texplus.com/texplus/comp5.html

[6] 문성빈 외, "적합성 피드백을 이용한 전문 검색 시스템의 검색 효율성 증진을 위한 연구", 정보 관리학회지, 제10권, 제2호, pp.43-67, 1993

[7] 박혁로 외 "효율적 정보검색 환경구현", pp.201-210, 연구개발정보센터, 1998

[8] Mckinim, E.J., Sievert, M., Johnson, E.D., & Mitchell, J.A., "The Medline/Full_Text Research Project", 42, pp297-307, Journal of the American Society for Information Science, 1990

[9] Blair, D.C. & Maron, M.E.,"An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System", Communication of the ACM, Vol28, No.3, pp.289-299, 1985