네비게이션 정보 추출에 의한 XML 본문검색시스템

XML Full-Text Retrieval System by Extracting Navigation Information

강남규(Nam-Gyu Kang)*, 이응봉(Eung-Bong Lee)**, 이석형(Seok-Hyoung Lee)*

초 록

최근, 키워드 기반의 정보검색의 한계를 극복하기 위한 구조문서 기반의 연구가 활발하게 진행되고 있지만, 실제 적용에는 많은 어려움이 존재한다. 본 고에서는 구조문서에 대한 본문검색시스템을 제안한다. 본문검색시스템에 적용된 문서는 XML로 구축된 국가연구개발 보고서를 대상으로 하였으며, XML 연구보고서의 DTD, 본문 간의 이동을 위한 네비게이션 정보추출, 본문검색을 위한 검색엔진의 적용 방안에 관하여 살펴본다. 본 시스템은 XML 문서에 대해 문서의 구조정보를 저장하고 이를 검색하여 다양한 형태로 열람할 수 있는 검색 엔진의 부재 상황을 본문검색이라는 방법으로 극복하기 위한 것이다.

ABSTRACT

Recently, to overcome the limit of keyword based retrieval system, the study based structured document has been studied. But it is hard for structured retrieval system to adapt a real service, in this paper, we propose a method of retrieval mechanism for the ful-text of XML documents. We explain DTD of XML based report, extracting navigation information and planing to adapt the retrieval system for article retrieval. Using the fulltext retrieval scheme, suggested system can be an alternative plan of professional structured based retrieval system.

키워드 : 구조문서, 구조문서검색, 네비게이션 정보, 본문검색, XML, structured document, structured retrieval system, navigation information, fulltext retrieval, KRISTAL

* 한국과학기술정보연구원 정보시스템연구실 (ngkang@kisti.re.kr, skyi@kisti.re.kr)

** 충남대학교 사회과학대학 문헌정보학과 조교수(eblee@cnu.ac.kr)

■ 논문 접수일 : 2002년 8월 일

■ 게재 확정일 : 2002년 8월 일

1. 서론

구조문서 기반의 정보검색에 대한 연구는 기존의 문서에서 제공하던 키워드 기반의 내용정보검색의 한계를 극복하고자 활발히 진행되고 있다. 기존의 문서와 달리 구조문서는 하나의 문서에 내용정보와 구조정보를 모두 포함하기 때문에 키워드 기반의 내용정보검색뿐만 아니라 문서의 논리적 구조정보에 대한 검색도 제공해야 한다. 이를 위해 XML 문서구조 정보 표현 방법, 색인 구조 및 구조정보검색 알고리즘들이 연구되고 있다.

하지만, 구조검색을 위해서는 구조문서 저장을 위한 저장시스템 구현, 트리 구조의 엘리먼트에 접근하기 위해 복잡한 연산의 수행, 트리 구조의 깊이에 따라 엘리먼트 정보를 표현하기 위하여 무한대의 저장공간 확보 등의 많은 어려움이 있다.

본 고에서는 구조문서 검색시스템의 구현에 따른 어려움을 극복하고자 구조문서에 대한 본문검색시스템을 제안한다. 본문검색시스템은 구조화된 문서를 대상으로 구조정보를 추출하고 제목 및 내용을 분리하여 저장 및 검색할 수 있도록 한다. 또한 상위 문서로의 이동, 동레벨의 이전 혹은 다음 문서로의 이동, 그리고 하위 문서로의 이동에 필요한 구조정보 등을 추출하여 문서 이동에 사용될 탐색정보로 이용하도록 한다. 또한 제목과 내용을 기반으로 한 검색을 수행하며 검색 결과의 간략 보기에는 장, 절 단락 등의 트리 구조 형태로 화면을, 상세보기에서는 해당 장, 절 또는 단락 내용을 보여준다.

본 고의 주요 내용에 대한 기술순서는 다음과 같다. 2장에서는 전자문서를 표현하기 위한 SGML과 이의 응용인 HTML에 대해 살펴보며, SGML과 HTML의 장점을 수용한 XMl에 대해서 알아본다. 3장에서는 본문검색시스템에 사용된 XML 연구보고서의 DTD와 구조정보 및 탐색정보를 위한 네비게이션(navigation) 정보 추출에 대해 살펴본다. 그리고 본문검색을 위해 추출된 네비게이션 정보를 검색시스템의 적재 형식으로 변환하는 방법 및 검색시스템에 대해 알아본다.

2. 문서기술언어

2.1 SGML

SGML(Standard Generalized Markup Languge)은 범용으로 데이터를 마크업하기 위한 표준화된 방법으로 설계되어 문서의 내용이나 내용구조를 정의할 수 있어 문서구조를 기반으로 한 응용에 사용될 수 있다. 하지만, 너무 복잡하여 SGML 처리시스템 구축이 쉽지 않고, 인터넷을 기반으로 하지 않기 때문에 서비스 제공에 많은 문제점이 있다. 이에 SGML의 한 응용으로 WWW(World Wide Web)상의 문서를 기술하기 위한 언어인 HTML(Hyper-Text Markup Language)은 플랫폼에 관계없이 사용이 가능하여 많이 이용되고 있다.

2.2 HTML

HTML은 SGML의 응용으로 SGML에서 하나의 DTD(Document Type Definition)를 갖는 SGML 문서로 볼 수 있다. HTML은 다운로드가 쉽고 빠르며, 이식성 등 사용이 편리한 장점을 가지고 있지만, 사용할 수 있는 태그는 HTML이 지원하는 범위 안에서만 사용할 수 있는 고정된 태그만을 제공하고, 다양한 레이

[표 1] SGML, HTML, XML 비교

	SGML	HTML	XML
출력형식언어	DSSSL	CSS	XSL
링크	Hytime	HTML	XLL
스펙	복잡	단순	단순
태그사용	사용자 정의 가능, 무제한	사용자 정의 불가능, 제한적	사용자 정의 가능, SGML보다는 제한적
브라우저	전용 브라우저	웹 브라우저	웹 브라우저
데이터 공유	가능	불가능(단방향)	가능(양방향)
문서작성	복잡	간단 용이함, 논리구조 작성 어려움	작성용이
문서검색	정확한 검색 가능, 문서구조 검색 가능	효과적 검색 어려움	정확한 검색 가능, 문서구조 검색 가능
문서 재사용	가능	불가능	가능
응용분야	방대한 내용, 구조의 기술문서	단순 구조 서	방대한 내용, 구조의 기술문서, 웹상의 교환문서

아웃이 지원되지 않으며 문서 자체가 구조화되어 있지 않아 효과적인 검색, 재사용, 검증 등이 불편하다.

2.3 XML

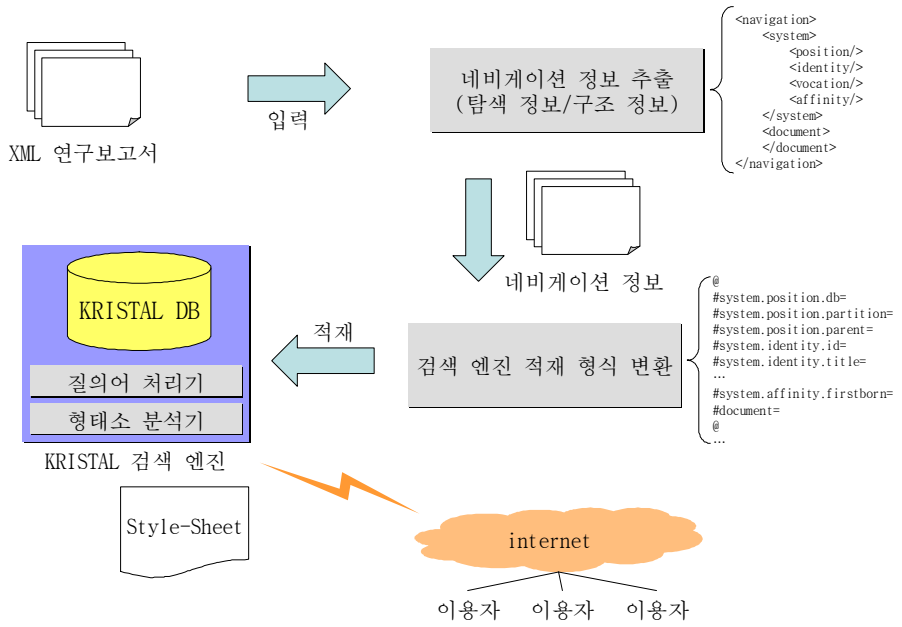
XML은 1996년 W3C에서 제안하여 1998년에 표준으로 제정된 것으로, 웹상에서 구조화된 문서를 전송 가능하도록 설계된 표준화된 텍스트 형식이다. HTML이 가지고 있는 장점인 웹상에서의 정보제공 능력과, SGML의 장점인 복잡한 문서구조를 제공, 이에 따른 문서 태그명도 자유롭게 제공한다는 측면에서 XML은 SGML과 HTML의 장점을 수용하고 단점을 보완하여 만들어진 웹문서 표준포맷이다. XML의 설계 목표는 인터넷상에서의 사용, 응용 프로그램의 폭 넓은 다양성 지원, SGML과의 호환, 용이한 문서처리 프로그램의 작성 등이라 할 수 있다.

또한 XML은 문서의 내용과 구조정보에 관한 데이터 기술을 위한 언어로써 구조화된 구조를 갖는 자기 서술적(self-describe)인 언어이기 때문에 응용에서 순수한 데이터로 사용될 수 있으며, 또한 효과적인 정보검색이 가능하다. 이것은 기존의 문서에서 제공하던 키워드를 기반으로 한 내용정보의 검색뿐만 아니라 문서의 논리적 구조정보에 대한 검색이 가능하다는 것을 의미한다.

3. XML 본문검색시스템

3.1 시스템 구성도

XML 본문검색시스템은 [그림 1]과 같이 XML 연구보고서에서 네비게이션 정보를 추출 모듈, 검색 엔진의 적재 형식으로 변환하는 모듈, KRISTAL 검색엔진으로 구성되어 있다.



[그림 1] XML 본문검색시스템

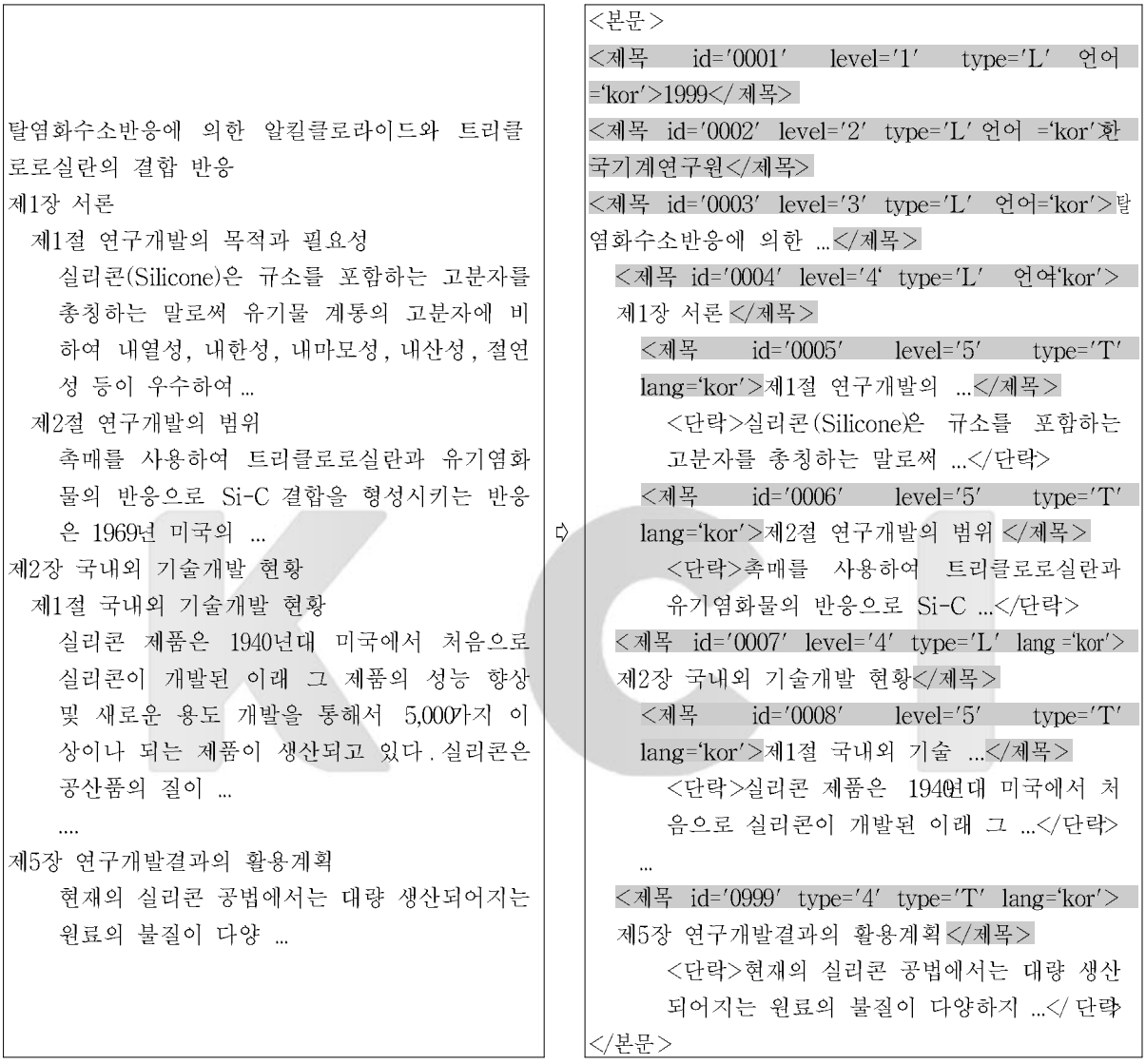
3.2 XML 연구보고서

본문검색시스템에 사용된 XML 연구보고서는 2001년 한국전산원의 "과학기술 문헌 및 산업기술기반 전문정보 DB 구축" 사업의 일환으로 구축되었다. 총 521건의 구축된 XML 연구보고서는 부, 요약문, 목차, 본문, 부록물, 부속물 등 6개의 구조로 구성되어 있으며, 본 본문검색시스템은 XML 연구보고서의 본문만을 대상으로 검색을 제공한다.

다른 문서들과 마찬가지로 연구보고서는 다양한 형태의 구조를 가지고 있다. 다양한 형태의 구조를 모두 수용하기 위하여, 구축된 연구보고서는 제목 엘리먼트와 그 속성만으로 문서의 모든 구조정보를 표현하였다.

[그림 2]는 연구보고서 원문을 연구보고서 DTD에 따라 태깅한 XML 문서의 예이다. 앞서 설명한 것과 같이 장, 절 그리고 하위의 부분의 구조정보는 새로운 엘리먼트로 정의하지 않고 하나의 제목 엘리먼트로 정의하여 단순한 형태로 표현을 하였고, 그 구조정보는 제목 엘리먼트의 속성에 표현하였다. 그리고 연구보고서의 년도, 연구수행기관 및 연구보고서의 제목 정보를 포함하기 위하여 3개의 제목 엘리먼트를 각각의 XML 문서작성시 추가하였다. [그림 2]에서는 제목 엘리먼트의 id 속성 값이 '0000', '0001'과 '0002'가 추가된 정보이다.

[표 2]는 제목 엘리먼트의 속성들을 나타낸 것으로 id, type, level, lang에 대한 설명이다. [표 2]에 나타난 바와 같이 제목 엘리먼트의 type 속성 값으로는 하위 문서의 존재 유·무를 파악할 수 있으며, level 속상 값으로는 현재 문서의 위치(depth)를 알 수 있다. 위 2개의 속성 값으로 전체 문서의 구조를 파악할 수 있으며, 이 값들은 향후 문서의 이동을 위한 탐색 정보로 이용된다.



[그림 2] XML 연구보고서

[표 2] 제목 엘리먼트의 속성

속성	설명
id	유일한 번호 표기 (XML 연구보고서의 목차에서 직접 링크될 수 있는 번호)
type	하위 제목이 있는 경우 'L'로 표기 하위 제목이 없는 경우 'T'로 표기
level	문서의 구조 정보(depth)를 표기(1, 2, 3, ...) '1'은 최상위를 뜻한다
lang	사용 언어에 따라 'kor' 또는 'eng'로 표기

[그림 3]은 XML 연구보고서의 DTD 중에서 본문과 제목 엘리먼트, 그리고 속성 정보만을 표기한 것이다. DTD로부터 본문은 1번 이상의 제목과 단락 또는 단락으로 구성됨을 알 수 있다. XML 연구보고서의 전체 DTD는 [부록 1]에서 살펴 볼 수 있다.

```
<!-- ELEMENT -->
<!ELEMENT 연구보고서 (부 ?, 전체요약서 ?), 요
약서, 목차, 본문, 부록물?, 부속물?))+>
<!ELEMENT 본문 ((제목?, (단락|삽화|...)*))+>
<!ELEMENT 제목 (#PCDATA|...)>
<!ELEMENT 단락 (#PCDATA|...)>

<!-- ATTLIST -->
<!ATTLIST 제목
    id ID #IMPLIED
    type CDATA #IMPLIED
    level NMTOKEN #IMPLIED
    lang CDATA #IMPLIED
>
```

[그림 3] XML 연구보고서 DTD

3.2 네비게이션 정보 추출

XML 본문검색시스템은 구조정보 표현을 위해 XML 연구보고서의 본문 엘리먼트 내용으로부터 네비게이션 정보를 추출하여 새로운 엘리먼트를 구성한다. 네비게이션 정보는 XML 연구보고서 본문의 제목 엘리먼트 단위로 구성되며 문서의 내용 및 구조정보를 모두 포함하기 위하여 [그림 4]와 같이 시스템(system)과 도큐먼트(document) 엘리먼트의 쌍으로 구성된다.

시스템 엘리먼트는 [표 3]과 같이 4개의 엘리먼트로 구성되는데, position 엘리먼트는 XML 본문검색을 위해 필요한 데이터베이스 정보를 기록한다. identity와 vocation 엘리먼트는 XML 연구보고서 제목 엘리먼트의 속성 정보를 추출하여 연구보고서의 구조정보를 표현할 수 있도록 한다. affinity 엘리먼트는 각 제목 엘리먼트간의 구조관계를 나타내며, 검색 결과에 대해 3방향 탐색(상위, 동 레벨의 이전, 다음)이 가능하게 한다. affinity 엘리먼트의 upper, previous, next, firstborn 속성 값들은

```
<본문>
<navigation>
  <system>
    <position dbase='' partition='' parent=''/>
    <identity id='' title=''/>
    <vocation type='' level=''/>
    <affinity upper='' previous='' next=''
    firstborn=''>
  </system>
  <document>
  ...
  </document>
</navigation>
  ...
</본문>
```

[그림 4] 네비게이션 정보

[표 3] 시스템 엘리먼트 및 속성 정보

엘리먼트	속성	설명
position	dbase	데이터베이스 명
	partition	데이터베이스 번호
	parent	상위 개념의 데이터 베이스 존재 유/무
identity	id	제목 엘리먼트 단위의 유일한 레코드 번호
	title	제목 엘리먼트의 내용
vocation	type	하위 제목이 있는 경우 'L'로 표기 하위 제목이 없는 경우 'T'로 표기
	level	문서의 구조정보(depth)를 표기 (1, 2, 3, ...)
affinity	upper	상위 레벨의 제목 레코드 번호
	previous	동 레벨의 이전 제목 레코드 번호
	next	동 레벨의 다음 제목 레코드 번호
	firstborn	하위 제목이 있는 경우 가장 첫 번째 나타난 제목의 레코드 번호

전체 문서 제목 엘리먼트들의 type과 level 값을 이용하여 추출하며 해당 제목의 레코드 번호를 가진다.

도큐먼트 엘리먼트는 XML 연구보고서의 장, 절에 해당하는 제목 엘리먼트의 시작부터 다음 제목 엘리먼트 사이의 내용을 추출하여 기록한다. 이러한 방식으로 추출된 도큐먼트의 내용에는 단락, 표, 그림, 수식 등 다양한 형태의 엘리먼트들이 존재한다. 그러나 이러한 엘리먼트들은 네비게이션 정보 추출과정에서 도큐먼트의 내용으로만 고려되며 본문 검색의 결과를 보여줄 때 style-sheet이 적용되어 화면에 보여진다.

```
<본문>
                    ...
<navigation>
<system>
  <position dbase='db' partition='1' parent='N'/>
  <identity id='0005' title='제1절 연구개발의 목적과 필요성'/>
  <vocation level='5' type='T'/>
  <affinity upper='0004' previous='N' next='0006' firstborn='N'/>
</system>
<document>
  <제목 id='0005' level='5' type='T' lang='kor'>
  제1절 연구개발의 목적과 필요성</제목>
  <단락>실리콘(Silicone)은 규소를 포함하는 고분자를 총칭하는 말로써 유기물 계통의 ...</단락>
</document>
</navigationt>
                    ...
</본문>
```

[그림 5] 네비게이션 정보 추출

[그림 5]는 XML 원시문서로부터 네비게이션 정보를 추출한 것으로 [그림 2]의 '제 1장 서론'의 '제 1절' 내용에 대한 네비게이션 정보이다. position 엘리먼트의 속성 값들은 시스템이 정해 놓은 값으로 결정되며 identity 및 vocation 엘리먼트의 속성 값은 제목 엘리먼트의 속성 값과 내용이 된다.

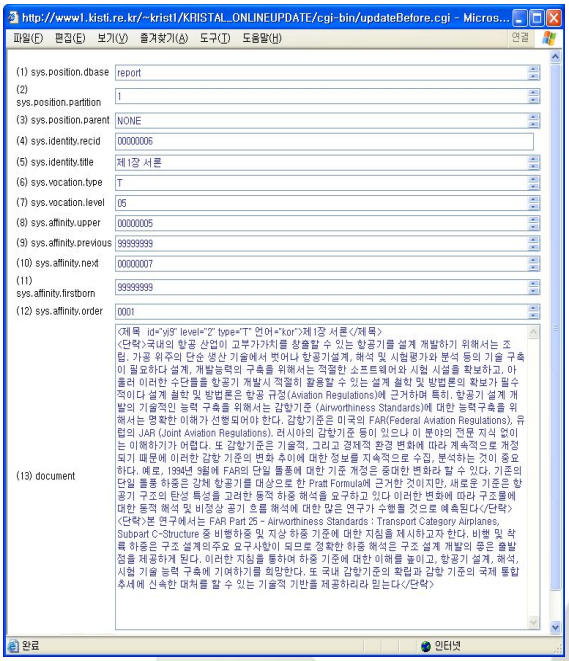
[표 4] affinity 엘리먼트 속성 정보

id	T	U	P	N	F
0001	1999	N	N	xxxx	0002
0002	한국기계연구원	0001	N	xxxx	0003
0003	탈염화수반응에 의한 알킬 ...	0002	N	xxxx	0004
0004	제1장 서론	0003	N	0007	0005
0005	제1절 연구개발의 목적과 필요성	0004	N	0006	N
0006	제2절 연구개발의 범위	0004	0005	N	N
0007	제2장 국내외 기술개발 현황	0003	0004	xxxx	0008
0008	제1절 국내외 기술개발 현황	0007	N	xxxx	N
...	...	...	...	...	...
0999	제5장 연구개발의 활용계획	0003	xxxx	N	N

[표 4]는 affinity 엘리먼트의 속성 정보에 대한 것으로 위와 같은 정보로 XML 연구보고서의 본문 구조정보를 한눈에 표현할 수 있으며, 하나의 네비게이션 정보로부터 상위 레벨, 동레벨의 이전 및 다음으로 이동할 수 있는 방법도 제공한다. [표 4]의 id, title(T)은 identity 엘리먼트의 속성 값이며 upper(U), previous(P), next(N), firstborn(F)은 affinity의 속성 값이다. 'N'으로 표기된 것은 해당 제목이 없다는 것을 의미하며 'xxxx'로 표기된 것은 해당 제목의 레코드 번호를 의미한다.

3.3 XML 본문검색

XML 본문검색을 위해 사용된 검색엔진은 한국과학기술정보연구원(KISTI)에서 자체 개발한 KRISTAL(Korea Research Information of Science & Technology Access Line) 버전 2.5를 사용하였다. 이 검색엔진은 한글/한자 변환



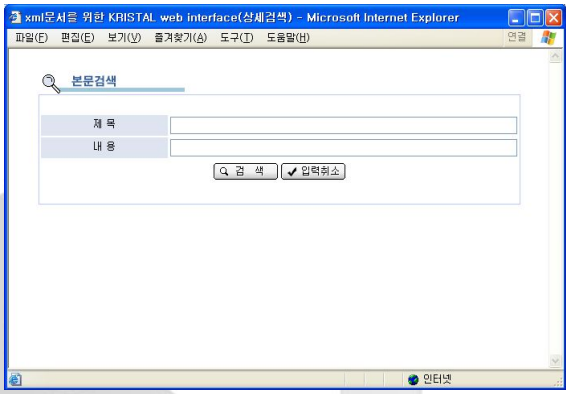
[그림 6] KRISTAL검색엔진 적재 파일 형식

색인 및 검색, 한글 형태소 분석, 복합 명사 분리 등의 완벽한 한글처리를 지원한다.

XML로 작성된 연구보고서에서 추출된 네비게이션 정보는 KRISTAL 검색엔진에 적재하기 위하여 텍스트 파일 형태로의 변환과정을 거치며, [그림 6]은 변환된 하나의 레코드를 예로 표현한 것이다. 변환된 파일은 레코드의 시작을 나타내는 시작 태그, 섹션(section : KRISTAL 검색 엔진에서 레코드를 구성하는 최소 단위의 문서 구성요소)의 시작을 의미하는 시작 태그, 섹션 이름, 섹션 값으로 구성된다.

섹션의 이름은 네비게이션 정보의 각 엘리먼트와 속성으로 만들어지는데, 예를 들어 identify의 id의 섹션 이름은 system.identify.id로 표현된다. 섹션 값은 시스템 엘리먼트의 경우, 속성 값으로 정의되고, 도큐먼트 엘리먼트는 엘리먼트의 내용이 섹션 값으로 정의된다. 본 시스템에서는 연구보고서의 제목과 단락 내용을 검색하며, 이 때 질의 대상이 되는 섹션은 [그림 6]의 system.identify.title과 document 섹션이 된다.

다음은 변환된 파일을 KRISTAL 검색엔진에 적재한 후, 검색 인터페이스와 검색된 결과를 출력하는 방법에 대해 설명한 것이다.



[그림 7] XML 본문검색 인터페이스

본문 검색 인터페이스는 [그림 7]과 같이 제목과 내용으로 검색할 수 있는 화면을 제공한다. [그림 7]에는 3.2절에서 설명한 것과 같이 모든 연구보고서의 문서 구조가 동일하지 않기 때문에 구조정보에 대한 검색 항목이 포함되어 있지 않다. 향후 연구보고서의 문서 구조가 동일한 형태로 구성된다면 위 검색 인터페이스 화면에 구조정보에 해당하는 레벨 정보를 제공하면 구조문서 검색도 가능하다.

본문 검색을 위하여 제목은 system.identity.title 섹션과 대응되고 본문의 내용은 document 섹션과 대응된다. 그리고 위에서 설명한 것과 같이 구조정보를 제공하기 위한 레벨 정보는 system.vocation.level 섹션과 대응된다.

제목과 본문내용에 대한 검색결과는 동일한 형태로 제공되며 제목 검색결과는 검색된 제목과 그 제목이 속한 상위 제목들이 제공된다.

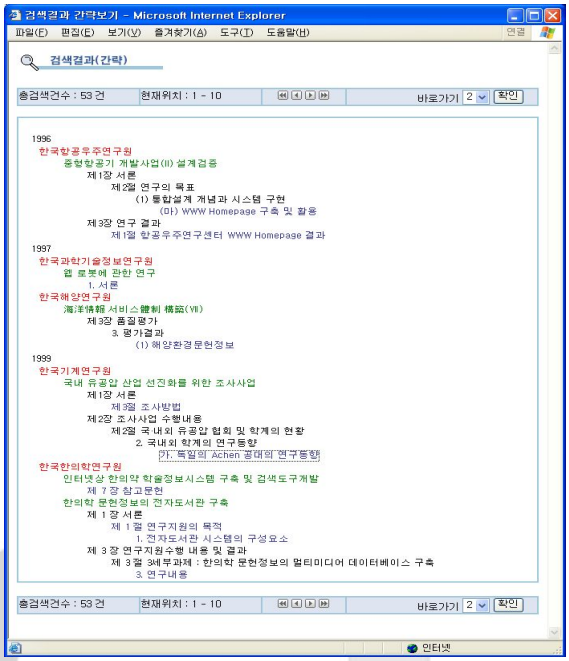
제목 검색결과와는 달리 본문내용 검색결과는 검색된 본문이 속한 제목과 상위 제목들이 제공된다.

검색의 결과는 간략보기와 상세보기 2개의 화면으로 출력된다. 간략보기는 검색된 결과를 트리 형태의 구조로 화면에 보여주는데 [그림 8]과 같은 방법으로 해당 레코드의 상위 레코드들을 찾아낸다.

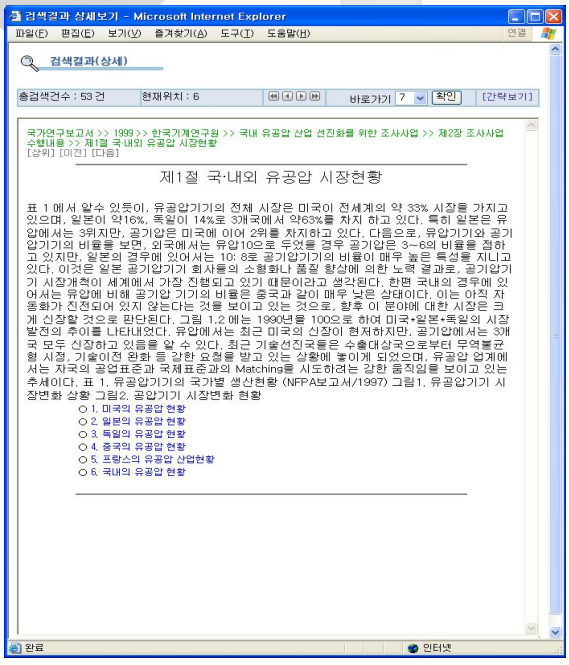
```
//
// id : 검색된 레코드의 system.identify.id
//
system = getSysInfo(id);
if(system.vocation.level < 2) return;

new tree[system.vocation.level]
level = system.vocation.level - 1;

tree[level --] = id;
do{
  system = getSysInfo(system.affinity.upper);
  tree[level --] = system.idenfity.id;
}while(system.affinity.upper != 'N')
```

[그림 8] 상위 레코드 찾기

[그림 8]의 방법은 다음과 같은 절차로 수행된다. 검색된 레코드의 시스템 정보 중에서 vocation 엘리먼트의 level 속성 값을 알아낸다. level 속성 값이 2 이상일 경우에는 상위 레코드가 존재한다는 것을 의미하며 affinity 엘리먼트의 upper 속성 값을 참조하여 상위 레코드의 시스템 항목을 얻어낸다. 위 과정은 upper 속성 값이 'N'일 때까지 반복 수행되며 위와 같은 방법으로 알아낸 레코드의 시스템 항목들은 화면에 표시할 때 [그림 9]과 같은 형태의 트리구조로 보여준다.



[그림 9] XML 본문검색(간략보기)



[그림 10] XML 본문검색(상세보기)

[그림 10]와 [그림 11]은 [그림 9]의 검색 결과에 대한 상세보기 화면과 XML 레코드를 보

```
<record>
<location>
    <ltitle>국가연구보고서 >> 1999 >> 한국기계연구원 >> 국내 유공압 산업 선진화를 위한 조사사업
       >> 제2장 조사사업 수행내용 >> 제1절 국·내외 유공압 시장현황</ltitle>
    <litem recid='00003714'>[상위]</litem>
    <litem recid='N'>[이전]</litem>
    <litem recid='00003722'>[다음]</litem>
</location>
<system>
    <position dbase='report' partition='1' parent='NONE'/>
    <identity recid='00003715' title='제1절 국·내외 유공압 시장현황'/>
    <vocation type='L' level='5'/>
    <affinity upper='00003714' previous='N' next='00003722'/>
</system>
<document>
    <제목 id='00003715' level='5' type='L' lang='kor'>제1절 국·내외 유공압 시장현황</제목>
    <단락>표1에서 알수 있듯이, 유공압기기의 전체 시장은 미국이 전세계의 약 33% ...</단락>
    <단락>...</단락>
    <children>
        <child recid='00003716'>1. 미국의 유공압 현황</child>
        ...
        <child recid='00003721'>6. 국내의 유공압 현황</child>
    </children>
</document>
</record>
```

[그림 11] XML 검색 레코드

여준다. [그림 10]의 화면은 [그림 11]의 XML 문서에 style-sheet를 적용하여 나타낸 것이며, KRISTAL 검색엔진은 검색된 레코드의 시스템 정보를 이용하여 탐색을 위한 다음과 같은 처리를 수행한다.

[그림 10]의 화면 상단에는 검색된 문서의 구조정보와 3방향 탐색(상위, 동 레벨의 이전, 다음)을 위한 링크가 제공된다. 3방향 탐색은 affinity 엘리먼트의 upper, previous, next의 속성 정보를 이용한다. 이 정보들은 [그림 11]의 location 엘리먼트로 표현된다. 그리고 화면 가운데는 본문의 내용이 기술되며, 검색된 문서

```
//
// id : 검색된 레코드의 system.identify.id
//
system = getSysInfo(id);
if(system.vocation.type == 'T') retrun;

new link[MAX];
system = getSysInfo(system.affinity.firstborn);
link[index ++] = system.identify.id;
do{
   system = getSystemInfo(system.affinity.next);
   link[index ++] = system.idenfity.id;
}while(system.affinity.next != 'N')
```

[그림 12] 하위 레코드 찾기

가 하위 문서를 가지고 있을 경우 하단에는 하위 문서로 가는 링크가 제공된다. 하위 문서로 가는 링크는 vocation 엘리먼트의 type 속성 값이 'L'일 경우에 제공되는데 [그림 11]의 children 엘리먼트로 표현되며 [그림 12] 와 같은 방법으로 하위 문서들을 찾아낼 수 있다.

4. 결론

본 고에서는 XML 문서를 대상으로 네비게이션 정보 추출에 의한 XML 본문검색시스템을 제안하였다. 본 시스템은 기존 서지정보 중심의 검색시스템과는 달리 구축된 XML 문서의 본문을 대상으로 검색을 수행하기 때문에 검색된 문서의 구조 및 본문의 내용을 한번에 제공한다는 큰 장점이 있다. 또한 제목과 내용정보 추출에 의한 단순한 구조정보 표현으로 구조정보 표현에 소요되는 비용의 절감 및 구조정보를 최대한 손상시키지 않고 데이터베이스를 구축 및 검색할 수 있다는 장점을 갖고 있다. 그리고 상위 문서로의 이동, 동 레벨의 이전, 다음 문서로의 이동, 그리고 하위 문서로의 이동에 필요한 구조정보 등을 추출하여 문서 이동을 위한 탐색정보로 활용하여 다양한 접근 방법을 제공한다. 그러나 이미 구축된 XML 문서의 구조정보 표현에 있어서 하나의 제목 엘리먼트와 속성정보로 구성되어 있지 않다면 본 시스템에 바로 적용하기 힘들다는 단점이 있다.

네비게이션 정보 추출에 의한 XML 본문검색시스템은 단순한 구조 정보 표현, 다양한 접근 방법 등의 많은 특징을 바탕으로 기존 서지정보 중심의 검색시스템과는 차별화된 서비스를 제공할 수 있을 것이며 또한 현재 연구 및 개발되고 있는 구조문서 검색시스템의 대안으로 활용될 수 있을 것이라 생각한다.

<참고문헌>

조윤기, 조정길, 이병렬, 구연설. 2001. "XML 문서에 포함된 구조 정보의 표현과 검색." *정보처리학회지*.

Brain Lowe, Justin Zobel, Ron Sacks- Davis. 1995. "A Formal Model for Databses of Structured Text." *Proceedings of the 4th International Conference on Database Systems for Advanced Applications*.

Toung Dao. 1998. "An Indexing Model for Structured Documents to Support Queries on Content, Structure and Attributes." *Proceedings of Advances in Digital Libraries '98*.

David Hunter 외 5인. 2000. "Beginning XML." 정보문화사.

정회경. 1999. "차세대 웹 문서 표준 XML." *정보처리학회지*.

W3C. 2002. "http://www.w3.org/", W3C(Word Wide Web consortium).

정보시스템연구실. 2002. "KRISTAL 매뉴얼," 한국과학기술정보연구원

[부록 1]

```
<?xml version="1.0" encoding="EUC-KR"?>

<!-- ##################################### -->
<!-- ## XML DTD For Research Report ## -->
<!-- ##################################### -->

<!ENTITY % doctype "연구보고서">
<!ENTITY % p.float "그림그룹 | 각주 | 주">
<!ENTITY % p.lib "헤더?, 저자명*, 표제사항, 형태사항?, 발행사항?, URL?, 기타사항">
<!ENTITY % p.el "삽화">
<!ENTITY % p.em.ph "윗첨자 | 아래첨자 | 폰트">
<!ENTITY % p.rf.ph "각주참조 | 주참조 | 그림참조 | 표참조 | 삽화참조 | 참고문헌참조 | 수식참조 | 제목참조 | 부참조">
<!ENTITY % p.lst.d "목록">
<!ENTITY % p.form "수식">
<!ENTITY % p.tbl "표">
<!ENTITY % p.zz "%p.el; | %p.tbl; | %p.lst.d; | %p.form;">
<!ENTITY % p.zz.ph "%p.em.ph; | %p.rf.ph;">
<!ENTITY % s.zz "단락|인용단락|참고문헌|%p.zz; |%p.float;">
<!ENTITY % m.sec "( 제목?, (%s.zz;)*)">
<!ENTITY % m.ph "(#PCDATA|%p.zz.ph;|%p.form;|%p.float;)*">
<!ENTITY % m.pseq "(단락, (단락|참고문헌|(%p.zz;))*)">
<!ENTITY % m.figheader "제목*">
<!ENTITY % m.date "#PCDATA">
<!ENTITY % m.fig "EMPTY">
<!ENTITY % m.toc "EMPTY">
<!ENTITY % a.id "id ID #IMPLIED">
<!ENTITY % a.rid "rid IDREF #REQUIRED
                  xml:link CDATA #FIXED 'simple'
                  xml:href CDATA #IMPLIED
                  role CDATA #IMPLIED
                  title CDATA #IMPLIED
                  inline (true|false) 'true'
                  show (embed|replace|new) 'replace'
                  actuate (onLoad|onRequest|other|none) 'onRequest' ">
<!ENTITY % a.sizes "SIZEX CDATA #IMPLIED
                  SIZEY CDATA #IMPLIED
                  UNIT CDATA #IMPLIED">
<!ENTITY % li.type "(solid | dot | dash | omit )">
<!ENTITY % h.align "(left | center | right)">
<!ENTITY % v.align "(top | middle | bottom)">

<!NOTATION tex PUBLIC "+//ISBN 0-201-13448-9::Knuth//NOTATION The TeXbook//EN">
<!NOTATION bmp PUBLIC "+//ISBN 0-7923-9432-1::Graphic Notation//NOTATION Microsoft Windows bitmap//EN">
<!NOTATION GIF PUBLIC "+//ISBN 0-7923-9432-1::Graphic Notation//NOTATION CompuServe Graphic Interchange Format//EN">
<!NOTATION jpg PUBLIC "+//ISBN 0-7923-9432-1::Graphic Notation//NOTATION Joint Photographic Experts Group raster//EN">
<!NOTATION tiff PUBLIC "+//ISBN 0-7923-9432-1::Graphic Notation //NOTATION Aldus/Microsoft Tagged Interchange File Format//EN">

<!ELEMENT %doctype; ((부?, 전체요약서?), (요약서, 목차, 본문, 부록물?, 부속물?))+>
<!ELEMENT 부 (#PCDATA)>
<!ELEMENT 전체요약서 (일반정보, 연구원정보+, 요약본문+, 키워드*)>
<!ELEMENT 요약서 (일반정보, 연구원정보, 요약본문+, 키워드*)>
<!ELEMENT 고유번호 (#PCDATA)>
<!ELEMENT 일반정보 (과제코드?, 고유번호, 발행일, 주관부처명, 사업명, 중점과제명, 세부과제명*, 보고서유형)>
<!ELEMENT 과제코드 (#PCDATA)>
<!ELEMENT 발행일 (%m.date;)>
<!ELEMENT 주관부처명 (#PCDATA)>
<!ELEMENT 사업명 (#PCDATA)>
<!ELEMENT 중점과제명 (#PCDATA)>
<!ELEMENT 세부과제명 (#PCDATA)>
<!ELEMENT 보고서유형 (#PCDATA)>
<!ELEMENT 연구원정보 (수행기관, 연구책임자?, 연구원*)>
<!ELEMENT 수행기관 (#PCDATA)>
<!ELEMENT 연구책임자 (성명+, 주민번호?, email?, 전화번호?)>
<!ELEMENT 주민번호 (#PCDATA)>
<!ELEMENT email (#PCDATA)>
<!ELEMENT 전화번호 (#PCDATA)>
<!ELEMENT 연구원 (직책, 성명, 담당업무?)>
<!ELEMENT 직책 (#PCDATA)>
<!ELEMENT 성명 (#PCDATA)>
<!ELEMENT 담당업무 (#PCDATA)>
<!ELEMENT 요약본문 (%m.sec;)+>
<!ELEMENT 제목 %m.ph;>
<!ELEMENT 단락 (#PCDATA | %p.zz.ph; | %p.zz; | %p.float;)*>
<!ELEMENT 인용단락 (#PCDATA | %p.zz.ph; | %p.zz; | %p.float;)*>
<!ELEMENT 키워드 (#PCDATA)>
<!ELEMENT 목차 %m.toc;>
<!ELEMENT 본문 (%m.sec;)+>
<!ELEMENT img %m.fig;>
<!ELEMENT 삽화 %m.fig;>
<!ELEMENT 목록 (헤더?, 항목)*>
<!ELEMENT 헤더 %m.ph;>
<!ELEMENT 항목 %m.pseq;>
<!ELEMENT 윗첨자 %m.ph;>
<!ELEMENT 아래첨자 %m.ph;>
<!ELEMENT 폰트 %m.ph;>
<!ELEMENT 각주참조 (#PCDATA)>
<!ELEMENT 주참조 (#PCDATA)>
<!ELEMENT 그림참조 (#PCDATA)>
<!ELEMENT 표참조 (#PCDATA)>
<!ELEMENT 삽화참조 (#PCDATA)>
<!ELEMENT 참고문헌참조 (#PCDATA)>
<!ELEMENT 수식참조 (#PCDATA)>
<!ELEMENT 제목참조 (#PCDATA)>
<!ELEMENT 부참조 (#PCDATA)>
<!ELEMENT 그림그룹 (%m.figheader;, (그림내용 | 서브그림+), (자료출처 | 기호설명*)>
<!ELEMENT 그림내용 %m.fig;>
<!ELEMENT 서브그림 (%m.figheader;, 그림내용)>
<!ELEMENT 자료출처 (#PCDATA)>
<!ELEMENT 기호설명 %m.ph;>
<!ELEMENT 각주 (%m.pseq;)>
<!ELEMENT 주 (%m.pseq;)>
<!ELEMENT 표 (제목*, 표두문?, (부표 | 표본문)+, (자료출처 | 기호설명*)>
<!ELEMENT 부표 (제목?, 표본문)>
<!ELEMENT 표본문 (img|행)+>
<!ELEMENT 행 (열)+>
<!ELEMENT 열 (셀)+>
<!ELEMENT 셀 %m.pseq;>
<!ELEMENT 표두문 (#PCDATA)>
<!ELEMENT 수식 (#PCDATA | img)*>
<!ELEMENT 부속물 (참고문헌)>
<!ELEMENT 참고문헌 (분야?, (%p.lib;)+)+>
<!ELEMENT 분야 (#PCDATA)>
<!ELEMENT 저자명 (#PCDATA)>
<!ELEMENT 표제사항 (기사명, 표제?)>
<!ELEMENT 기사명 (#PCDATA)>
<!ELEMENT 표제 (#PCDATA)>
<!ELEMENT 형태사항 (Vol | No | Part | Page)+>
<!ELEMENT Vol (#PCDATA)>
<!ELEMENT No (#PCDATA)>
<!ELEMENT Part (#PCDATA)>
<!ELEMENT Page (#PCDATA)>
<!ELEMENT 발행사항 (발행기관 | 출판지 | 출판사 | 출판년)+>
<!ELEMENT 발행기관 (#PCDATA)>
<!ELEMENT 출판지 (#PCDATA)>
<!ELEMENT 출판사 (#PCDATA)>
<!ELEMENT 출판년 (#PCDATA)>
<!ELEMENT URL (#PCDATA)>
<!ELEMENT 기타사항 (#PCDATA | img)*>
<!ELEMENT 부록물 (부록+)>
<!ELEMENT 부록 (%m.sec; | 요약본문)>

<!ATTLIST 저자명 언어 CDATA #IMPLIED>
<!ATTLIST 기사명 %a.id; 자료유형 CDATA #IMPLIED 본문언어 CDATA #IMPLIED>
<!ATTLIST 표제 ISSN CDATA #IMPLIED ISBN CDATA #IMPLIED>
<!ATTLIST URL DOI CDATA #IMPLIED>
<!ATTLIST 그림그룹 %a.id;>
<!ATTLIST 서브그림 %a.id;>
<!ATTLIST 각주 %a.id;>
<!ATTLIST 주 %a.id;>
<!ATTLIST 삽화 %a.id; %a.sizes; NAME ENTITY #IMPLIED TYPE (tex | bmp | gif | tiff) #IMPLIED SCALE CDATA "100">
<!ATTLIST 그림내용 %a.id; %a.sizes; NAME ENTITY #IMPLIED SCALE CDATA "100">
<!ATTLIST img %a.id; %a.sizes; NAME ENTITY #IMPLIED TYPE (tex | bmp | gif | tiff) #IMPLIED SCALE CDATA "100">
```

<!ATTLIST 폰트 FAMILY CDATA #IMPLIED SIZE CDATA #IMPLIED COLOR CDATA #IMPLIED POSTURE (normal | italic | oblique) #IMPLIED WEIGHT (light | medium | bold) #IMPLIED RULING (undline | strkthru | overbar) #IMPLIED STYLE CDATA #IMPLIED>
<!ATTLIST 부 %a.id;>
<!ATTLIST 부록 %a.id;>
<!ATTLIST 수식 %a.id; CONTENT ENTITY #IMPLIED>
<!ATTLIST 표 %a.id;>
<!ATTLIST 목록 %a.id;>
<!ATTLIST 각주참조 %a.rid; index CDATA #IMPLIED>
<!ATTLIST 주참조 %a.rid; index CDATA #IMPLIED>
<!ATTLIST 부참조 %a.rid;>
<!ATTLIST 그림참조 %a.rid;>
<!ATTLIST 표참조 %a.rid;>
<!ATTLIST 삽화참조 %a.rid;>
<!ATTLIST 참고문헌참조%a.rid;>
<!ATTLIST 제목참조 %a.rid;>
<!ATTLIST 수식참조 %a.rid;>
<!ATTLIST 일반정보 공개여부 (공개 | 비공개) #REQUIRED>
<!ATTLIST 제목 %a.id; type CDATA #IMPLIED level NMTOKEN #IMPLIED 언어 CDATA #IMPLIED>
<!ATTLIST 중점과제명 언어 CDATA #IMPLIED>
<!ATTLIST 세부과제명 언어 CDATA #IMPLIED>
<!ATTLIST 연구원정보 유형 CDATA #IMPLIED>
<!ATTLIST 성명 언어 CDATA #IMPLIED>
<!ATTLIST 요약본문 언어 CDATA #IMPLIED>
<!ATTLIST 키워드 언어 CDATA #IMPLIED>
<!ATTLIST 행 ho.align %h.align; #IMPLIED ve.align %v.align; #IMPLIED line.type %li.type; #IMPLIED>
<!ATTLIST 열 colno CDATA #IMPLIED rowno CDATA #IMPLIED ho.align %h.align; #IMPLIED ve.align %v.align; #IMPLIED line.type %li.type; #IMPLIED celltype (1 | 2 | none) #IMPLIED>
<!ATTLIST %p.lst.d; type %li.type; #IMPLIED>