濟州大學校 尖端技術硏究所 論文集, 第13卷, 2號, pp.99-104, 2002. 12
Cheju Nat'l Univ. Res. Inst. Adv. Tech. Jour., Vol.13, No.2, pp.99-104, 2002. 12

제목 내의 키워드 역할정보를 이용한 연구 논문 검색 시스템

김 영 민* · 이 상 준**

Efficient Research Paper Searching System using the Keyword Role in the Title

Young-Min Kim* and Sang-Joon Lee**

ABSTRACT

Many information searching systems support the keyword based search using database. However, the results are becoming larger day by day. So we need some methods such as semantic-based search or keyword role-based search to minimize the result volume and to help users to get more correct informations which they want. In this paper, we propose the efficient paper search system using the role informations found in the paper title. For this system, we extracted the role meaing words between the keywords from the paper title. And we stored these role informations and the keywords as XML document within 6 elements. Finally we could seach the papers efficiently using the role informations represented by XML.

Key Words : Paper search, Information search, XML

Ⅰ. 서 론

데이터베이스와 웹의 발달과 더불어 정보 저장 활동이 활발하게 이루어지고 있다. 특히 최근에는 수많은 학회가 새로이 생겨나고 있고 대학, 연구소 등에서 수많은 연구논문들을 발표하고 있어서 이에 대한 체계적 정보 관리가 필요한 실정이다. 현재 많은 학술정보 제공 DB들이 있으며 이들에 대한 검색 서비스가 제공되고 있지만 저장데이터베이스의 대형화로 인해 점차 불필요한 검색결과까지 제공하게 되는 일이 허다해지고 있다. 이에 좀더 논문 검색자의 의도에 맞는 검색서비스 제공이 중요한 이슈로 대두되고 있다.

현재 정보 검색의 방법으로는 문서 생성자에 의해 미리 정의된 키워드를 대상으로 검색하는 방법, 정보 데이터 내에서 주제어 및 주제문장을 찾아내게 하고 이를 통해 문서를 검색하는 주성분 분석방법을 이용하는 방법[1], 질의에 사용된 단어와 연관된 단어들까지 질의어에 포함시켜 검색하는 질의어

* 제주대학교 컴퓨터공학과 대학원
Department of Computer Engineering, Cheju Nat'l Univ.
** 제주대학교 통신.컴퓨터공학부, 첨단기술연구소
Faculty of Computer Engineering, & Prod., Engineering, Res. Inst. Adv. Tech., Cheju Nat'l Univ.

확장 검색방법. 질의에 사용되는 단어들의 시소러스를 구성하여 연관단어를 알아내는 방법[2.3.4.5]. 키팩트라는 문서의 주된 내용을 대표하는 {중심어, 종속어} 쌍을 추출하고 이중 가장 빈도수가 많은 키팩트들에 가중치를 두어 이용자의 질의와 검색 대상의 유사도를 비교하는 방법등이 연구되고 있다[6.7]. 또한 의미망 구축후 의미망을 통해 용어간의 연관성 측정. 개념적으로 관련된 용어를 추론. 검색에 반영하는 방법도 있으며[8]. 검색어의 의미를 이용한 검색을 수행하기 위해 질의어에 대한 형태소 분석을 통해 의미어를 추출하고. 의미어로 재구성된 템플릿을 사용하여 의미기반 검색을 하기도 한다[7.9]. 또한 최근에는 XML의 구조정보를 이용하여 자신의 원하는 정보를 담고있는 XML정보를 검색해 오는 시맨틱웹(Semantic Web)에 관한 연구도 이루어지고 있다[10.11.12].

이러한 수많은 연구에도 불구하고 연구 논문 검색과 같이 내용검색이 아닌 논문 요약 정보 및 문서 검색이 주된 경우 대부분 키워드와 제목. 요약내의 문자열과 질의어와의 유사성만을 이용한 정보검색이 이루어지고 있어서 검색자가 원하는 정확한 범주의 결과만을 제시하지 못하고 있다. 이들의 경우 대부분 이용자가 원하던 정확한 정보를 검색하는 것이 아니라 이용자가 원하는 정보를 포함하고 있는 방대를 정보를 제시할 뿐이며. 이용자로 하여금 제시된 결과내에서 또 다시 검색을 수행하게 하는 이중 작업을 요구한다. 또한 키워드 생성에 있어서도 문서 생성자가 정의하고 있어서 정확하고. 획일적인 표현을 사용하지 않는 경우 검색의 정확성을 더욱 떨어뜨리게 된다.

본 논문에서는 이러한 검색 결과의 정확성을 개선하기 위하여 연구 논문들의 제목내에서 나타나는 키워드와 키워드들의 역할정보를 이용하여 검색자가 원하는 논문만을 검색하는 방법을 제안하고 있다. 이를 위해 데이터 저장시 제목 내에서 키워드와 역할어를 추출하고 이를 XML형태의 메타정보로 구성한 후 함께 저장되게 하였고. 검색할 때는 키워드. 제목 문자열과 함께 메타정보를 사용하여 검색하게 하므로 사용자 의도가 반영될 수 있는 검색을 수행하게 하였다.

II. 제안 검색 시스템의 설계

본 장에서는 논문 제목내에서의 역할어 추출 및 검색 방법. 그리고 제안하는 검색시스템의 구조에 대해 설명한다.

Fig. 1에서 보듯이 키워드 기반의 OR 검색은 전체원에 해당하는 검색 결과를 보이고. And 검색은 큰 사각형 영역과 같은 검색결과의 크기를 갖는다. 하지만 본 논문에서 제안하는 메타정보 검색은 그 안에 포함되어 있는 색상을 갖는 사각형처럼 부분집합을 형성한다는 것을 예측 할 수 있다. 즉 단순 키워드 기반의 검색보다는 역할정보를 이용한 검색 결과는 그 부분집합일 수 밖에 없으며 최적의 경우 가장 검색 자가 의도하는 정확한 결과를 제시할 수 있게 된다. 그리고 최악의 경우도 기존 키워드 기반 검색과 동일한 경우일 뿐이라는 것을 알 수 있다.

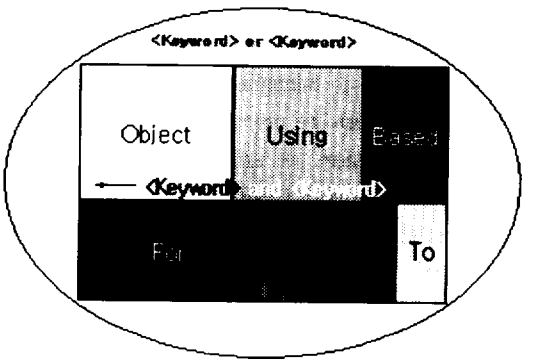

Fig. 1. Searched spaces for keyword based searching and meta information based searching

2.1 키워드 역할 정보 추출

본 논문이 검색 대상으로 하는 연구 논문들은 논문 제목에 있어서 일정한 패턴이 나타난다. 그 패턴을 유심히 살펴보면 키워드들은 일정한 역할을 맡고 있다는 것을 분석할 수 있다. 어떤 키워드들은 제목 내에서 목적으로 사용되고. 연구의 결과로 표현되기도 하며. 때로는 연구 결과를 이끌어 내기 위해 사용된 기술 및 배경 이론을 의미하기도 한다. 이들의 일반적인 패턴을 모형화 하면 Table 1과 같이 나타낼 수 있다.

Table 1. Patterns of keyword role in the research paper title

역할별 키워드종류	역할 의미
In 키워드	연구의 대상. 범위를 나타냄
Based 키워드	기반 이론 및 기술을 나타냄
Using 키워드	사용 도구 및 방법론을 나타냄
For 키워드	연구의 목적 등을 나타냄
To 키워드	연구 방향을 나타냄
Object 키워드	연구 결과물을 나타냄

이러한 가정 하에 연구 논문의 제목을 대상으로 각각의 키워드의 역할을 의미하는 어휘들을 추출하여 '역할어'라 명명한다면, Table 2와 같이 키워드별 역할어를 정의할 수 있다.

Table 2. Keyword roles and role-implying-words

keyword roles	words which imply a role in the title
In	에서 /
Based	기반 / 갖는 / 기초한 / 기반한 / 포함한
Using	이용한 / 통한 / 고려한
For	위한 / 지원하는 / 용 / 관한 / 지향하는
To	로 /
Object	needless

2.2. 키워드 역할에 대한 메타정보 생성

Table 2와 같이 추출된 역할어를 이용하여 논문제목에 대한 메타정보인 키워드 역할정보를 나타내는 XML문서를 생성한다. Table 2에서 추출된 역할어를 중심으로 논문제목을 분할하고, 각각의 분할된 키워드들을 역할어를 의미하는 Table 3의 6가지의 XML 엘리먼트를 이용하여 표현하면 Table 4와 같이 각 제목들에 대한 메타정보를 생성할 수 있다.

Table 3. XML document structure for paper title

```
〈subject〉
  〈title〉 원 제목 〈/title〉
  〈object〉 object역할 키워드 〈/object〉
  〈based〉 based 역할 키워드 〈/based〉
  〈using〉using역할 키워드 〈/using〉
  〈for〉 for 역할 키워드 〈/for〉
  〈in〉 in 역할 키워드 〈/in〉
  〈to〉 to 역할 키워드〈/to〉
〈/subject〉
```

Table 4. Research paper titles transformed into XML form

〈subject〉〈object〉테스트 데이타 자동생성 도구 : AUTEG 〈/object〉 〈based〉UML〈/based〉〈/subject〉
〈subject〉 〈object〉HTML 문서 XML 자동 변환 〈/object〉 〈based〉유사 패턴〈/based〉〈/subject〉
〈subject〉〈object〉XML 문서 분석 기법〈/object〉〈based〉유사성〈/based〉〈/subject〉
〈subject〉〈object〉스키마 추출방법〈/object〉〈using〉엘리먼트 정보〈/using〉〈in〉XML 문서〈/in〉〈/subject〉
〈subject〉〈object〉컴포넌트 저장소 검색〈/object〉〈based〉확장된 소프트웨어 컴포넌트 서술자에〈/based〉〈/subject〉
〈subject〉〈object〉수학식 편집 및 표현 시스템〈/object〉〈based〉MathML에〈/based〉〈/subject〉
〈subject〉〈object〉상품 카탈로그 설계 및 적용〈/object〉〈based〉XML〈/based〉〈/subject〉
〈subject〉〈object〉XML Gateway 설계 및 구현〈/object〉〈for〉Legacy 데이터베이스〈/for〉〈/subject〉
〈subject〉〈object〉목표문서 인식기법〈/object〉 〈based〉XML〈/based〉〈for〉EDMS〈/for〉〈/subject〉
〈subject〉〈object〉웹 컨텐츠 관리 시스템 설계 및 구현〈/object〉 〈based〉XML〈/based〉〈/subject〉
〈subject〉〈object〉스키마 추출〈/object〉〈using〉 엘리먼트 정보〈/using〉〈in〉XML 문서〈/in〉〈/subject〉
〈subject〉〈object〉효율적 질 처리〈/object〉〈using〉분할 저장시중복〈/using〉〈to〉XML 데이터 RDBMS〈/to〉〈/subject〉
〈subject〉〈object〉XML 문서 저장 및 검색 시스템〈/object〉〈using〉관계형 데이터베이스와 XQuery 〈/using〉〈/subject〉

2.3. XML 메타정보를 이용한 논문 검색

XML구조의 문서에 대한 저장 및 검색방법에 대한 많은 연구가 이루어지고 있다. Table 5는 그중에 XPath를 이용한 XML 노드 검색 예를 보여주고 있다.

Table 5. Query examples for XML node search using XPath

XPath Query	Meaning
/paper[//title ~ ='xml*']	<title>엘리먼트중 'xml'로 시작하는 노드 검색
/paper[//object ~ ='*검색*']	<object>엘리먼트중 '검색'을 포함하는 노드 검색
/paper[//object ~ ='*편집*' and //for ~ ='*xml*']	<object>가 '검색'을 포함하고, <for>에 'xml'을 포함할때
/paper[title ~ ='*에이전트*' or //in ~ ='*db*']	<in>에 'db'를 포함하거나, <title>이 '에이전트'를 포함
/paper[@ino:id > 2000]/title	ino:id 속성값이 2000보다 큰 모든 <title>엘리먼트 출력

2.4. 시스템의 구조

제안하는 시스템은 Fig. 2와 같이 구성된다.

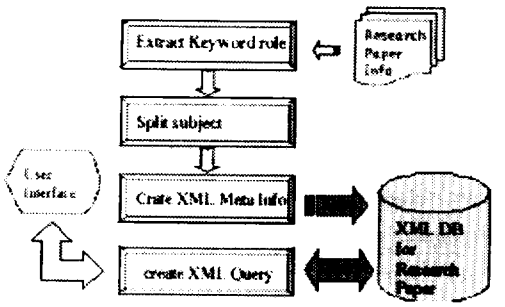

Fig. 2. System Architecture.

이 시스템은 3가지 주요 구성요소로 XML기반 논문정보 저장 시스템, 논문정보 중 제목내 키워드 역할 분석 및 XML메타정보 생성 모듈, 사용자 질의에 대한 XML Xpath기반 쿼리 생성기이다. 본 논문에서 논문정보 저장시스템으로 Tamino XML Server를 사용해 XML을 저장하고 Xpath를 이용하여 쿼리할 수 있도록 하였다[13,14].

III. 구현 및 결과

이 장에서는 앞에서 제시한 시스템을 구현하고 테스트한 결과를 기술한다.

3.1 구현 환경

본 논문은 Windows2000 운영체제상에서 Tamino XML DB를 논문정보 저장 시스템으로 사용하였고, http://cseric.cau.ac.kr 에서 키워드 기반 검색을 이용해 각 xml관련 논문 546개, 에이전트 관련 논문 520개, 무선관련 논문 539개, 인식 관련 논문 805개, 보안 관련 논문 666개를 Java Servlet을 이용해 검색하였으며, 역할정보 추출을 위해서는 스트링 처리가 비교적 자유로운 ASP기반으로 수행하였고, 이를 다시 리눅스 환경에서 Java DOM API를 이용해 XML문서로 생성 후 XML저장시스템인 Tamino Server에 입력 후 검색에 사용하였다.

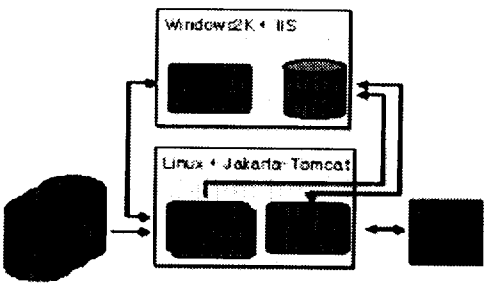

Fig. 3. Implementation Environment.

3.2. 역할정보 추출 결과

중복된 논문 정보를 제외한 총 2705개의 한글 논문들에 대해 패턴 분석을 수행한 후 각각에 대해서 XML형태의 메타정보를 생성하였으며 아래 Fig. 4는 변환 결과 화면을 보여준다.

2705개의 논문 제목을 대상으로 역할어 추출을 수행하였을 때 각 제목 내에서 역할어들이 검출된 횟수는 Table 6과 같았다.

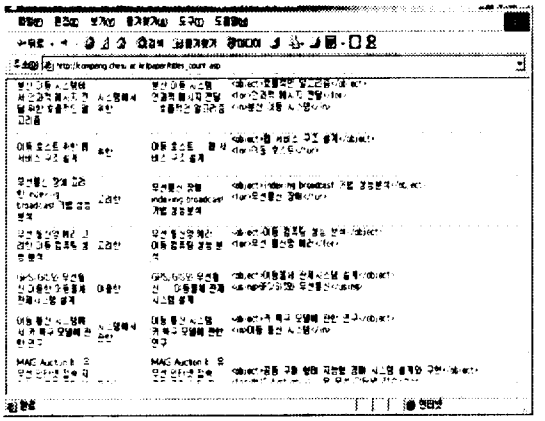

Fig. 4. XML based meta information creation.

Table 6. Frequencies for each role in whole titles

object	based	for	using	in	to
2705	674	781	883	63	500
100%	24%	29%	33%	2%	19%

3.3 검색 결과

다음 Fig. 5는 사용자로부터 검색조건을 입력받기 위한 사용자 인터페이스 화면이다. Fig. 6은 검색결과인 XML문서를 보여준다. Fig. 7은 Fig. 5에 XSL을 적용하여 HTML로 표현한 결과이다. Table 7과

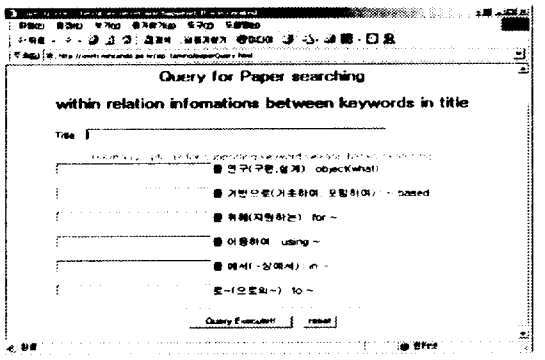

Fig. 5. User Interface for paper search.

Table 7. Average counts of searched informations

keyword-and	keyword-or	XML meta-info
40.40	532.20	16.20

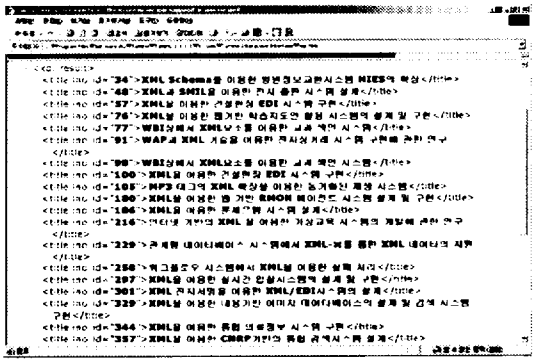

Fig. 6. Searched XML meta informations.

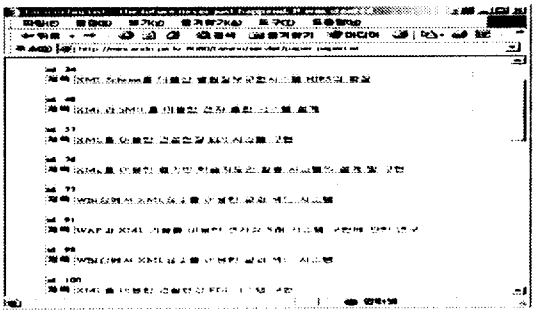

Fig. 7. Searched Research paper informations.

Table 8. Counts in keyword based search and XML meta-info based search

target words	method	used query	count
시스템 + xml	using keyword	시스템 and xml	127
		시스템 or xml	975
	using XML meta-info	시스템 using xml	26
		시스템 for xml	8
에이전트 + 보안	using keyword	에이전트 and 보안	19
		에이전트 or 보안	520
	using XML meta-info	에이전트 for 보안	2
		에이전트 object 보안	13
무선 + 변환	using keyword	무선 and 변환	11
		무선 or 변환	335
	using XML meta-info	무선 object 변환	10
		무선 for 변환	1
문서 + 인식	using keyword	문서 and 인식	6
		문서 or 인식	370
	using XML meta-info	문서 for 인식	1
		문서 object 인식	5
xml + 검색	using keyword	xml and 검색	39
		xml or 검색	461
	using XML meta-info	xml for 검색	15
		xml using 검색	0

Table 8에서는 역할어 메타정보를 이용한 검색과 키워드 기반 검색의 효율성 차이를 보이고 있다.

Ⅵ. 결 론

결론적으로 키워드기반의 학술정보 시스템은 그 결과가 방대하여 사용자로 하여금 추가적인 검색을 수행하게 하는 단점을 안고 있으며 사용자 의도를 정확히 반영해 주지 못하는 실정이다. 여기에 논문 제목 속에 숨겨진 역할정보를 추출하여 이를 메타정보로 활용한다면 사용자가 의도하는 정확한 대상 논문을 검색할 수 있을 뿐 아니라 그 결과 집합의 수도 최적의 수로 줄어든다는 것을 알 수 있었다. 향후에는 공학 연구논문 이외에 기타 논문들까지도 확장하기 위한 개념 및 역할어 연구가 필요하고 6가지의 패턴 정보 뿐만이 아닌 다양한 패턴정보도 연구해볼 필요가 있으며, 이러한 역할어 기반 검색방법을 디지털 도서관 및 웹 자원 검색에도 활용할 가치가 있을 것으로 여겨진다.

참고문헌

1) 이창범, 김민수, 백장선, 이귀상, 박혁로. 2001. 주성분분석을 이용한 주제어 기반 문서 자동 요약. KDBC(SIGDB-KISS) 2001 1호. pp.0191-0195.

2) 김영천, 이재훈, 문유미, 박병권, 이성주. 2001. 정보 검색에서 용어 가중치 재부여를 이용한 성능 증진에 관한 연구. 퍼지 및 지능시스템학회 논문지. 11권 9호. pp.0811-0816.

3) 김주연, 김병만. 2000. 용어 발생 유사도와 퍼지 추론을 이용한 질의 용어 확장 및 가중치 재산정. 한국정보과학회 논문지. B 27권 09호. pp.0961-0972.

4) 우선미, 유춘식, 김용성. 2001. 용어 연관성 분석을 이용한 사용자 위주의 문서순위결정 기법. 한국정보과학회 논문지. B 28권 2호. pp.0149-0156.

5) 양승원, 노희영, 2000. 시소러스를 이용한 XML 태그 검색 시스템. 정보과학회 2000년 추계학술대회. 27권 2호. pp.0145-0147.

6) 김수희, 남효돈. 2000. 정보검색에서 정확률의 향상을 위해 키팩트의 가중치 부여. 한국정보과학회 논문지. D 27권 4호. pp.0627-0636.

7) 장명길, 김현진, 장문수, 최재훈, 오효정, 이충희, 허정. 2001. 의미기반 정보검색. 한국정보과학회지. 19권 10호. pp.0007-0018.

8) 노영희. 2001. 개념기반 검색 방법론. 2001국제컨퍼런스 IT21-정보처리학회. pp.0079-0110.

9) 박현규, 오종훈, 김명호, 최기선, 이광형. 2001. 퍼지 추론에 의한 자연언어 정보 검색. 한국정보처리학회 논문지. B 11권 9호. pp.0811-0816.

10) 김노환, 정충교. 2001. XML DOM을 이용한 웹문서 검색 알고리즘. 컴퓨터산업교육기술학회 논문지. 2권 6호. pp.0775-0782.

11) XML : eXtensible markup Language. http://www.w3.org/ XML

12) Semantic Web. http://www.w3.org/2001/sw/

13) XML Query. http://www.w3.org/XML/Query

14) Tamino XML Server. http://www.softwareag.com/tamino/