웹 정보의 자동 의미연계를 통한 학술정보서비스의 확대 방안 연구

A Exploratory Study on the Expansion of Academic Information Services Based on Automatic Semantic Linking Between Academic Web Resources and Information Services

정도헌* · 유소영** · 김환민*** · 김혜선**** · 김용광***** · 한희준******
Do-Heon Jeong · So-Young Yu · Hwan-Min Kim · Hye-Sun Kim · Yong-Kwang Kim · Hee-Jun Han

차 례

1. 서 론
2. 이론적 배경
3. 연구방법
4. 자동분류 및 검색결과 평가
5. 결론
• 참고문헌

초 록

이 연구에서는 KISTI NDSL의 학술논문 정보를 웹 학술자원과 연계하는 실험적 연구를 수행함으로써 KISTI의 정보 유통 서비스의 확대 가능성을 살펴보고자 하였다. 이를 위해 웹 학술자원을 수집하여 STEAK 시스템을 이용한 자동 의미 연계를 생성하고 이를 학술논문 검색결과와 결합하였다. 시스템의 검색 정확률을 평가한 결과 매크로 정확률은 62.6%, 마이크로 정확률은 66.9%를 보였으며, 자동연계 성능에 대한 전문가 평가는 76.7점을 보였다. 주제 범주별 전문가 평가는 본 연구를 통해 의미연계를 잘 수행하는 경우에 높게 측정되어 시스템적 성능과 동일한 경향을 보였다. 이 연구는 다양한 웹 학술자원의 서비스 연계를 위하여 논문정보로부터 생성한 언어자원을 의미색인에 사용한 것으로 이를 통해 지속적인 웹 자원의 학술적 활용에 대한 가능성을 제시하고자 하였다.

키 워 드

데이터 해석, 웹 학술정보, 자동색인, 자동분류, 의미색인

* 한국과학기술정보연구원 정보서비스실 선임연구원
(Senior Researcher, Dept. of Information Service, KISTI, heon@kisti.re.kr)
** 연세대학교 대학원 문헌정보학과 박사과정
(Ph.D. Candidate, Dept. of Library and Information Science, Yonsei University, sweety798@yonsei.ac.kr)
*** 한국과학기술정보연구원 정보서비스실 선임연구원
(Senior Researcher, Dept. of Information Service, KISTI, mrkim@kisti.re.kr)
**** 한국과학기술정보연구원 정보서비스실 선임연구원
(Senior Researcher, Dept. of Information Service, KISTI, hskim@kisti.re.kr)
***** 연세대학교 대학원 문헌정보학과 박사과정
(Ph.D. Candidate, Dept. of Library and Information Science, Yonsei University, ykkim@yonsei.ac.kr)
****** 한국과학기술정보연구원 정보기술연구실 연구원
(Researcher, Dept. of Information Technology Research, KISTI, hhj@kisti.re.kr)
• 논문접수일자: 2009년 2월 19일
• 게재확정일자: 2009년 3월 24일

ABSTRACT

In this study, we link informal Web resources to KISTI NDSL's collections using automatic semantic indexing and tagging to examine the possibility of the service which recommends related documents using the similarity between KISTI's formal information resources and informal web resources. We collect and index Web resources and make automatic semantic linking through STEAK with KISTI's collections for NDSL retrieval. The macro precision which shows retrieval precision per a subject category is 62.6% and the micro precision which shows retrieval precision per a query is 66.9%. The experts' evaluation score is 76.7. This study shows the possibility of semantic linking NDSL retrieval results with Web information resources and expanding information services' coverage to informal information resources.

KEYWORDS

Data Analysis, Web Academic Resource, Automatic Classification, Semantic Indexing

1. 서론

연구자들은 다양한 공식·비공식 학술집단을 통해서 공통의 관심분야 정보를 교류해오고 있다. 공식적인 학술집단은 그 결과물을 논문, 프로시딩과 같은 구조화된 학술자원의 형태로 발표하고 있다. 따라서 공식적 학술집단의 연구결과물은 비교적 용이하게 파악·접근할 수 있다.

이와 다르게 비공식 학술집단은 서신, 이메일, 전화, 구두 등 외부에서 접근이 어려운 형태로 학술 커뮤니케이션을 진행해 왔기 때문에, 그 성과물 역시 파악하기가 매우 어려웠다. 그러나 근래에 와서 인터넷을 활용한 사회연결망 서비스가 크게 확대·강화되어 웹 2.0 시대의 블로그 및 집단지성(Collective Intelligence)을 위한 다양한 서비스들은 불특정 다수를 대상으로 사실을 기록하고 자신의 의견을 표현할 수 있다는 점에서 이전의 공식 학술집단 및 학술자원과는 다른 양상을 지니고 있다. 뿐만 아니라 사회연결망 서비스를 통한 다양한 인적 교류 네트워크의 형성은 이러한 자원들이 유통될 수 있는 기반을 제공한다고 할 수 있다.

이는 기존의 구조화된 학술 정보원을 제공하는 전문 정보센터 및 도서관에서도 주목할 만한 변화로 볼 수 있다. 즉, 블로그 포스팅을 준전문 학술자원 또는 반구조화 전문 학술자원으로 이용할 수 있는 가능성을 시사하는 것으로 볼 수 있다.

최근 연구에 따르면 넷 세대(Net Gen)의 학생들은 Google과 같은 유명한 접근점을 통한 도서관 소장 정보의 접근에 익숙하며 이러한 서비스가 확대되길 바라고 있다(Griffiths and Brophy 2005; Lippincott 2005). 이들은 간단하고 시각적인 방식으로 주제 검색이 제공되길 바라며 구애받지 않고 주제 안내를 받을 수 있고, 공개된 웹 정보원과 도서관 정보서비스를 결합한 검색을 원한다. 따라서 도서관은 블로그를 학생들이 수업의 과제 등을 통해서 찾은 유용한 정보를 교환할 수 있는 수단으로써 개발할 필요가 있다고 지적하고 있다(Lippincott 2005).

이밖에도 연구자들의 집단 및 국가 내 연구소들 간의 비공식적 사회연결망에 의한 관계가 연구 성과를 향상시켜준다는 연구결과들이 지속적으로 나오고 있다는 점은 시사하는 바가 크다(Liebeskind et al. 1997; Baum et al. 2000; Owen-Smith et al. 2002).

이와 같은 일련의 경향들을 통해 볼 때 학술정보 유통에 있어서도 자원의 유형에 관계없이 연구자의 참여를 지원할 수 있고 자유로운 의견 교환 및 평가가 가능한 콘텐트를 확보하고 있다면, 연구자들에게 자원으로서의 유용성을 인정받을 가능성을 가질 수 있다는 것을 추측해 볼 수 있다.

이 연구에서는 NDSL을 통해 제공되는 과학기술 학술논문정보와 블로그, 커뮤니티를 통한 웹상의 학술정보 자원의 연계서비스 개발을 위하여 기계적 의미매칭을 사용하여 웹정보자원과 학술정보 자원을 연결하는 시스템을 제안하고자 한다. 그리고 실제 검색 및 의미색인 시스템을 이용한 실험적 연구를 수행함으로써 이 연구의 방법을 통한 웹상의 다양한 학술정보 자원의 수집 및 유통 가능성을 검증해보고자 한다.

2. 이론적 배경

2.1 주제 게이트웨이

웹 자원의 학술 자원 이용은 사실상 전자형태의 학술 정보원을 이용하는 것이 대부분이라고 볼 수 있다. 윤정옥(2003)의 연구에서 보는 바와 같이 웹 자원의 서지레코드를 생성한 경우를 국내외 학술 정보원에서 찾아보기가 어려웠다. 그러나 최근에 들어서는 웹 자원의 범위가 확장되면서 보고서 및 리포트, 블로그, 서지모음(citeulike) 등 다양한 정보원들을 학술적 웹 자원으로서 바라보고자 하는 시각이 확대되고 있다고 볼 수 있다. 다양한 웹 자원 중 현재 주목받고 있는 것은 블로그와 사회연결망 서비스에서 생산된 자원들이다.

이렇게 다양한 웹 자원이 생산됨으로써 이전에 상업적 크롤러가 접근하지 못했던 학술 데이터베이스를 지칭하였던 심층 웹(hidden Web, Bergman 2001)을 넘어서 표층 웹(visible web)상의 다양한 정보원들에 대한 관심이 도서관 및 학술정보서비스 측면에서 이

루어졌다.

대학도서관 및 학술정보센터도 90년대 후반부터 웹 자원을 자관의 자원에 포함시키거나 자관의 리포지터리를 작성하여 이용자의 정보 요구를 만족시키고자 하는 시도들이 있었으며, 이는 주제 게이트웨이 서비스로 발전하였다.

주제 게이트웨이(subject gateway)는 시스템적으로 정보자원 발굴을 지원하는 인터넷 서비스로써 주로 인터넷을 통해 접근할 수 있는 문헌, 개체, 사이트, 서비스 등의 자원으로의 링크를 제공하며 주제 구조를 통해 정보자원을 탐색하는 것을 그 특징으로 하고 있다(Koch 2000). 또한 이는 DESIRE 시스템에서 사용한 '주제기반 정보 게이트웨이(subject-based information gateway, SBIG)'라는 용어와 동의어로 볼 수 있다(Koch 2000).

주제 게이트웨이로는 INFOMINE(http://infomine.ucr.edu), BUBL(http://bubl.ac.uk/), GEM(http://www.thegateway.org/), Intute(http://www.intute.ac.uk/) 등이 있고 국내에서도 서울대학교 도서관 및 숙명여자대학교 도서관 등에서 웹 자원을 포함하는 주제 게이트웨이 서비스를 하고 있다.

INFOMINE은 웹 자원 형태의 대학 수준의 연구 및 교육 자료를 인터넷으로 접근할 수 있는 주제 게이트웨이이다. 이 서비스는 1994년 미국 University of California, Riverside 도서관의 프로젝트로서 시작되었다. 여기에는 사서들이 생성한 2만6,000개의 링크와 크롤러를 통해 생성한 7만5,000여개의 링크를 포함하고 있으며 전자저널, 교과서, 프로시딩 등

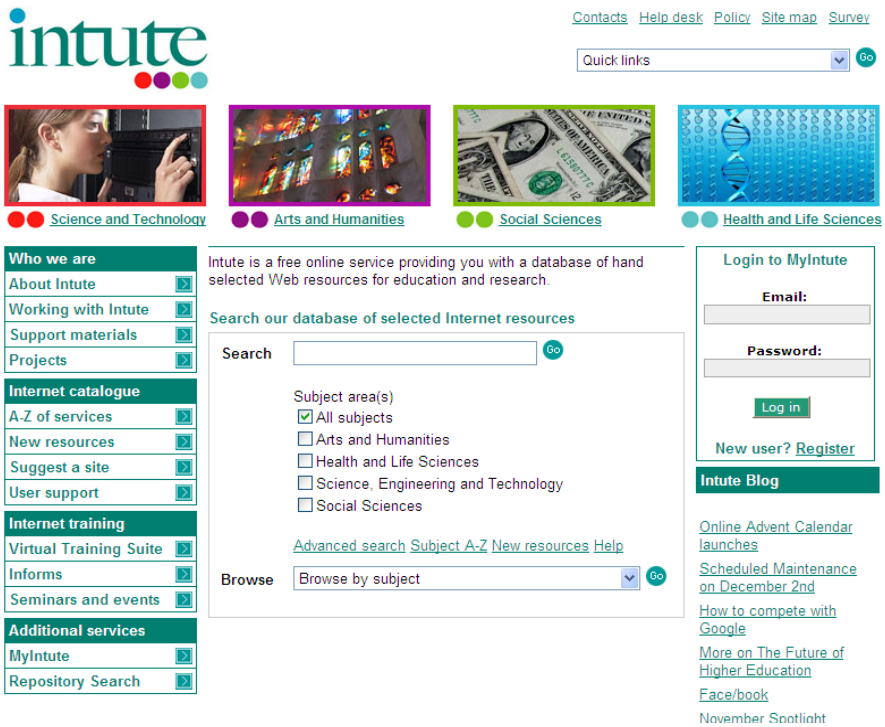
〈그림 1〉 INFOMINE (http://infomine.ucr.edu/)

의 자원을 포함한다. 전체 5만6,000여 개의 자료를 포함하고 있다.

주제 분야별로는 생물학·농학 및 의학, 경영경제, 문화학, e-Journal, 정부정보, 지도·GIS, 물리학·공학·전산 및 수학, 사회과학, 인문학, 시각 및 행위 예술의 자료를 가지고 있다. 자원의 유형별로 살펴보면 전자저널 및 연속간행물, 뉴스레터, 이미지 등의 데이터베이스, 디렉터리, 가상 도서관 검색 엔진, 참고정보원, 교육 자료 및 매뉴얼, 교과서, 커리큘럼, 단행본 및 텍스트 자원, 지도, eprint 및 preprint 자료 등을 포함한다.

이 시스템은 2개의 계층 구조를 가지고 있어서, 첫 번째 계층에 있는 원 자원을 두 번째 계층에서 자동적으로 선택하여 기술 자원을 작성하는 방식을 취하고 있다. 현재 브라우징, 검색, 제한 검색, 고급 검색 기능을 제공한다.

Intute는 웹상의 교육 및 연구 자원의 접근 가능성을 높이고 평가 및 협업을 통해 양질의 자원을 제공하고자 하는 데 그 목적을 두고 있는 사이트이다. 현재 과학·기술, 인문학, 사회과학, 생명과학·의학의 네 개 주제 집단으로 크게 나뉘어 65개 주제의 11만3,900개 이상의 데이터를 제공하고 있다. 이들은 'Joint Information Systems Committee (JISC)'의 재정지원을 받고 있으며 특히 인문학 분야



〈그림 2〉 Intute (http://www.intute.ac.uk/)

는 'Arts and Humanities Research Council (AHRC)'의 지원을 추가로 받고 있다. 또 Intute Consortium은 University of Manchester, Heriot-Watt University, University of Oxford을 포함한 7개 대학으로 구성되었으며 이 외에도 70여 개의 기관의 주제전문가들이 참여하고 있다. Intute 사이트에서는 브라우징 및 검색 외에도 개인화 서비스(My Intute), 세미나, RSS 등의 서비스를 제공한다.

그러나 이러한 주제 게이트웨이 서비스들은 현재 다양한 형태의 학술정보 자원을 수집하여 분류 및 조직을 통해 제공하고 있지만, 학술정보원과 웹 자원을 의미적으로 분류하거나 매칭하는 서비스는 제공하고 있지 않다. 웹 자원은 형식화된 학술정보원과는 달리 키워드가 구조화되어 있지 않기 때문에 단순한 자동화 시스템으로는 분류하는 데 한계가 있으며, 전문가의 수작업만으로 분류하기에는 자원의 양이 너무 방대하여 현실적인 서비스의 어려움이 따르게 된다. 따라서 정보서비스를 위한 웹 자원의 지속적인 발굴과 유지를 위해서는 단순한 웹 자원 수집 이상의 의미적 연결 및 서비스 제공을 위한 노력이 필요하다고 할 수 있다.

2.2 언어자원을 이용한 의미색인

언어자원에 기반한 의미색인에 관한 연구들은 특히 웹 2.0 기반의 폭소노미(Folksonomy)와 소셜 태깅(Social Tagging)의 출현으로 풍부한 웹 자원의 언어자원 및 의미색인의 필요성에 따라 더욱 부각되었다고 할 수 있다.

폭소노미의 네트워크의 특성을 분석하거나 (Cattuto et al. 2007), 이용자가 웹상에서 태깅한 정보를 분석하여 동시발생한 키워드 간에 의미관계를 생성하는 개념적 모델을 제안한 연구들이 있다(Cattuto, Loreto and Pietronero 2006). Cattuto 등(2006)의 연구에서는 다양한 소셜 태깅 서비스들에 나타난 이용자의 태깅 행태를 분석하여 그 안에서 동시 출현하는 색인어들이 매우 정확한 수준에서 일치한다는 것을 측정하였다. 또 Cattuto 등(2007)은 후속 연구를 통하여 폭소노미의 네트워크적 특성을 분석하였다. 이 연구에서는 Del.icio.us와 BibSonomy 데이터를 이용하여 부여된 태그의 간단한 통계적 분석부터 동시 출현한 태그들을 이용해 네트워크적 속성인 경로거리(path length), 최근접 이웃강도(nearest neighbor strength) 등을 분석하여 이용자의 태깅 행태를 보다 자세히 분석하고자 하였다.

이는 폭소노미를 이용자들이 생성한 개념구조로 바라보는 연구들로 태깅어들은 의미색인어로 볼 수 있다. 즉, 웹상의 언어 자원들은 다양한 이용자들에 의해서 형성된 하나의 의미적 색인어의 네트워크로 볼 수 있으며 출현한 개별적 키워드만으로는 내용 해석이 불충분 할 수 있음을 고려해야 한다고 할 수 있다.

웹상의 의미색인의 행태를 분석하고자 하는 연구들과 함께 지식체계 전반의 색인어들 간의 관계를 이용하여 문헌들의 의미적으로 표

현하고 분류하고자 하는 연구들이 이루어졌다 (Moschitti and Basili 2004; Bloehdorn and Hotho 2007; 정도헌, 최희윤 2007). Moschitti과 Basili(2004)는 문헌 자동분류 실험에서 문헌 표현을 위한 자연어 처리 기법으로 출현 키워드, 구문 정보, WordNet의 synset을 이용한 의미색인어 등을 사용하여 이들 간의 성능을 비교하였다. Bloehdorn과 Hotho (2007)도 문헌 자동분류에 Wordnet, MeSH와 같은 기존의 학술적인 자원을 통해 구조화된 지식 표현인 의미망 색인어를 이용하여 문헌을 표현하고 분류하고자 하였다. 또 정도헌과 최희윤(2007)은 다국어 자동의미색인 시스템인 STEAK를 제안하고 이를 통해 문헌의 자동분류 성능을 향상시키고자 하였다.

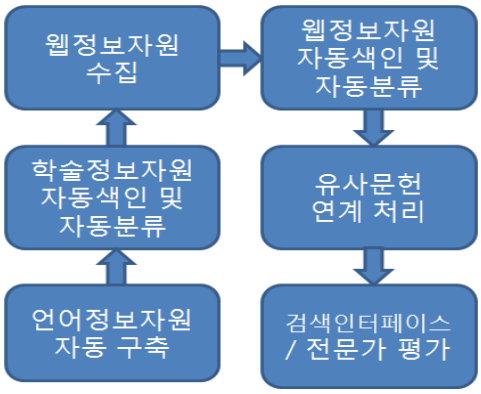
정도헌 등(2007)은 STEAK 시스템을 이용한 응용 연구에서 과학기술 전 분야의 전문용어를 추출하여 주제별 전문성을 측정하고, 자동분류성능 변화와의 연관성을 비교하였다. 연구결과 학술논문의 자동분류 시 키워드의 매칭률보다는 전문용어의 전문성이 성능에 영향을 주는 것을 확인하였다. 특히 이 연구에서는 저자들이 사용한 비통제 키워드를 이용하여 실험한 결과, 기계, 식품, 의약학 분야 등의 주제 분야에서 저자 키워드의 전문성이 높고 자동분류 성능이 우수한 것을 확인하였다.

이러한 연구들은 지적 체계로 구조화되어 있는 WordNet, MeSH 등의 언어자원 또는 STEAK와 같이 기계적으로 자동 생성한 어휘망을 이용하여 문헌들을 표현하고 분류하고자 한 시도들이다. 이들은 학술적 언어자원을 통해 의미색인을 시도했다는 점에서 웹상의 언어자원인 폭소노미를 의미색인의 한 방식으로 바라보고자 하는 연구들과 일맥상통한다고 할 수 있다. 특히 상호 이질적인 성격을 갖는 학술 자원과 웹 자원 간의 연계를 위해서는 각 자원을 표현하고 있는 색인어의 특성을 파악하고 상호 의미적인 연계를 하기 위한 자동화된 기법을 개발할 필요가 있다.

3. 연구 방법

3.1 연구 개요

이 연구의 개요는 〈그림 3〉과 같다. 먼저 KISTI가 보유하고 있는 학술논문, 연구보고서, 특허, 산업표준 등의 NDSL 데이터를 색인한 후 이를 이용하여 STEAK(S&T Ter minology system for the Evaluation and Analysis of Knowledge) (정도헌 등 2006; 정도헌 등 2007)의 의미색인 시스템 및 의미해석 시스템의 학습에 이용하였다. 그리고 웹 상의 정보자원 중 블로그와 오픈코스웨어, 기관 리포지터리 등을 중심으로 몇 개의 주요 사이트를 선정하여 웹 데이터베이스를 구축하였다. 이렇게 선정된 웹사이트로부터 자체 개발한 크롤러를 이용하여 웹 자원을 수집하고 KISTI 정보자원과 같은 방식으로 STEAK를 이용한 의미색인을 수행하였다. 크롤러는 Beautiful Soup과



〈그림 3〉 연구 개요

같은 Python HTML/ XML parser를 이용하고 히스토리 관리 및 크롤링의 깊이 단계를 조절하는 등의 기능을 추가하여 개발하였다.

시스템의 성능 평가는 자동분류의 성능 및 검색 성능으로 평가하였다. 자동분류의 성능은 중복 분류를 허용하여 분류 정확률을 측정하였다. 검색 성능은 검색 시 의미색인을 통해 학술논문 중 유사한 문헌과 연결되어 있는 웹 자원을 유사문헌으로 제공하게 하여 시스템 정확률과 전문가 검색결과 평가를 측정하였다. 시스템 정확률은 검색결과로 제시된 웹 문헌 중 상위 10개에 대하여 적합/부적합을 전문가 집단이 평가하도록 하였다. 전문가 검색결과 평가는 주제 분야별 전문가가 직접 작성한 질의를 검색하여 나온 유사 웹 문헌에 대해 리커트 3점 척도로 평가하도록 하였다.

3.2 웹 학술자원의 선정 및 수집

웹 학술자원 수집 후보 리스트 작성을 위하여 기존 연구들을 분석한 결과 대부분은 현재 업데이트가 지속적으로 이루어지지 않는 사이트들이 많았다. 또한 이 연구에서 이용하고자 하는 블로그 및 사회연결망 서비스를 제공하는 사이트를 학술 정보원으로 확인한 연구들은 발견하기 어려웠다. 따라서 KISTI의 주제분류 중 과학기술 전 분야에 해당하는 주제 범주명 및 STEAK 색인어를 이용하여 Google BlogSearch를 수행하여 검색결과 상위 100개 사이트를 각 검색어별로 수집하였다. 후보 사이트들을 직접 방문하여 사이트의 내용을 확인하고 검색된 사이트 내의 외부 링크에서 이 연구에서 사용하고자 하는 학술적 정보원을 가지고 있는 사이트들도 추가하였다. 검색 및 수집 사이트 선정은 2008년 5월부터 2008년 6월까지 이루어졌다.

수집 대상은 인터넷을 통해 전자형태로 제공되는 웹 학술자원 모두를 대상으로 하였다. 이때 수집 과정에서 확인된 기관 리포지터리, 오픈액세스, 온라인저널은 '유형I 웹 학술자원'으

로, 상용 학술데이터베이스 등에 비해 접근이 자유롭고 최신성 및 유용성을 가진 최신동향, 뉴스그룹, 온라인 포럼, 이용자 참여 백과사전 사이트 등을 '유형II 웹 학술자원'으로 보았다. 그리고 기존의 구조화된 학술자원으로 볼 수 있는 수업자료를 웹상에서 제공하는 OCW를 별도로 보았다. 이러한 웹 학술자원의 유형에 대한 정의는 지극히 조작적인 것으로 후속 연구를 통해 더 일반화할 수 있을 것으로 생각된다.

유형I 자원은 기관 리포지터리, 오픈액세스, 온라인저널로 분류하였다. 기관 리포지터리는 유수 대학 및 정부 기관에서 제공하는 학술 정보원으로 각 기관에서 구독하거나 생산 수집한 자원들을 수집 제공하는 사이트로 정의하였다. 그리고 본 연구에서 정의한 오픈액세스는 연구자 버전의 연구 성과물을 아카이빙한 Golden OA를 말한다. 따라서 이들은 원문제공에 제한이 없다. 반면에 온라인저널은 출판사 사이트에서 제공되는 출판사 버전의 연구성과물들로 대부분 최신호에 대해서는 접근권한을 제공하지 않고 있다. 또는 온라인저널 형태로만 제공되는 경우도 있다. 온라인저널의 경우에는 상업적으로 유통되는 학술지와 중복되는 부분이 발생할 수 있으나, 이 연구는 웹 학술자원의 유효성 또는 접근 가능성에 대한 실험적 연구이므로 이러한 정보자원을 포함하여 연구를 진행하였다.

유형II 자원은 유형I 자원에 비해 세부 유형이 비교적 다양하고, 구분하는 데 어려움이 있었다. 본 연구에서 사용한 유형II 자원은 대체적으로 최신동향 사이트가 많았다. 그 외에도 뉴스그룹, 온라인 포럼 사이트, 이용자 참여 백과사전 사이트도 수집 대상에 포함시켰다.

검색포털의 경우 사이트 자체 내 페이지보다는 공학 및 과학기술 분야의 여러 블로그의 내용을 검색할 수 있게 해주고 접근 링크를 제공해주는 일종의 포털, 또는 메타 블로그 사이트이다. 뉴스그룹은 소속 구성원에 의해 주기적으로 소식을 올리고 의견을 주고받는 사이트이며, 이용자 참여 백과사전 사이트는 Wikipedia와 유사한 형태로 운영되는 지구과학 분야 사이트를 분류한 것으로 전문 내용을 작성하면 이를 peer-review에 의해 검증하는 방식으로 되어 있었다.

그리고 유명한 국외 블로그 사이트의 하위 디렉토리 중 과학기술 및 교육 분야에 위치한 블로그들도 수집대상으로 삼았다. 이는 Google BlogSearch 검색결과 내에서 각 블로그 사이트의 페이지들이 다수 출현한 것을 확인하고 이들의 내용을 검토해 본 결과 학술자원으로써의 가능성을 가지고 있다고 판단되었기 때문에 포함시킨 것이다. 여러 최신동향 사이트들은 대부분 .com 도메인을 사용하고 있는 사이트들로 업계의 최신동향을 블로그 형식으로 빠르게 업데이트하는 형태를 취하고 있었다.

〈표 1〉 유형 I 웹 학술자원 수집 후보 사이트

세부유형	사이트	URL
기관 리포지터리	Connexions	http://cnx.org
	eScholarship Repository	http://repositories.cdlib.org/escholarship/
	Highwire Press	http://highwire.stanford.edu
	Pudue e-Pubs	http://docs.lib.purdue.edu
	ResearchNow	http://researchnow.bepress.com
	ScholarlyCommons @Penn	http://repository.upenn.edu
	ISTE	http://www.iste.org
	SLAC	http://www.slac.stanford.edu
	Tree of Life Web Project	http://www.tolweb.org/tree/
	US Forest Service	http://www.fs.fed.us
	ICSU(International Council for Science)	http://www.icsu.org
오픈액세스	arxiv.org	http://arxiv.org
온라인저널	Cardiovascular Ultrasound	http://cardiovascularultrasound.com
	SCitation	http://scitation.aip.org
	CSHprotocols	http://cshprotocols.cshlp.org
	PharmSciTech	http://aapspharmscitech.org
	AJP	http://ajprenal.physiology.org
	Cambridge Journal Online	http://journals.cambridge.org
	Diabetes Care	http://care.diabetesjournals.org
	European Heart Journal	http://eurheartj.oxfordjournals.org
	Heart online	http://heart.bmj.com
	IMA	http://imamat.oxfordjournals.org
	IOVS	http://www.iovs.org
	JACC	http://content.onlinejacc.org
	JAP	http://jap.physiology.org
	JBC	http://www.jbc.org
	JCS	http://jcs.biologists.org
	Protein Science	http://www.proteinscience.org
	The Plant Cell	http://www.plantcell.org
	WJES	http://www.wjes.org

이는 기존의 학술정보원보다 비교적 최신성 있는 정보를 제공하는 것으로 이 연구에서 확인된 새로운 형태의 자원으로 생각된다.

가상도서관(virtual library) 사이트는 유형 I 웹 자원의 기관 리포지터리와는 차이가 있는 것으로 구분하였다. 왜냐하면 이 유형의 사이트들은 기관에서 생산한 연구 성과 뿐 아니라 연구용 데이터, 최신 동향 등의 다양한 정보원을 함께 제공하는 포털사이트들이기 때문이다. 따라서 이들은 자관 생산 및 자관이 구독하고 있는 구조화된 학술정보를 제공하는 기관 리포지터리와는 차이가 있다고 할 수 있다. 이는 주제전문 게이트웨이와도 유사하지만 그 범위 및 정보자원의 질이 보다 다양하며 실제적으로 후속 연구를 통하여 확보해야 할 유형 II 웹 자원 사이트 유형의 하나라고 생각된다. 이 때 이러한 사이트 자원의 authority는 추후 고민해야 할 부분이 될 것이다. 커뮤니티 사이트의 경우에는 다양한 형태의 학술 자원을 제공한다고 볼 수 있는데 블로그, 포럼 등의 형태가 모두 존재하고 있다. 또 온라인 포럼 기능만을 주로 제공하는 사이트들은 포럼이라는 세부유형으로 분류하였다. 유형II 웹 자원의 수집대상 사이트는 〈표 2〉와 같다.

〈표 2〉 유형II 웹 학술자원 수집 후보 사이트

세부유형	사이트	URL
검색포털	Globalspec	http://engineering-tools.globalspec.com/TechLib/
뉴스그룹	newsgroups.derkeiler.com	http://newsgroups.derkeiler.com/Archive/Sci/
	tech-archive.net	http://sci.tech-archive.net
이용자참여 백과사전	Encyclopedia of Earth	http://www.eoearth.org
블로그	Livejournal 일부	http://www.livejournal.com/technology
	Wordpress 일부	http://news.wordpress.com/category/science/
	About.Com 일부	http://www.about.com/education/
	About.Com 일부	http://www.about.com/compute/
	About.Com 일부	http://www.about.com/health/
	Scienceblogs	http://scienceblogs.com
	Scientific Blogging	http://scientificblogging.com
	Wired Science	http://blog.wired.com/wiredscience/
	Arxivmath'Journal	http://arxivmath.livejournal.com
	The Math Mojo Chronicles	http://mathmojo.com/chronicles/
	Blogtica.com	http://www.blogtica.com
	Real Time Economics	http://blogs.wsj.com/economics/
	Let's Play Math!	http://letsplaymath.wordpress.com/
	LogicMatters	http://logicmatters.blogspot.com/
	Marginal Revolution	http://www.marginalrevolution.com/
	Math-Blog	http://math-blog.com/
	Mathematics Weblog	http://www.sixthform.info/maths/

	MathNotations	http://mathnotations.blogspot.com/
	MATLAB Central	http://blogs.mathworks.com/loren
	The Big Picture	http://bigpicture.typepad.com/
	Wild About Math!	http://wildaboutmath.com/
	Best-Health-Report	http://www.best-health-report.com/
	Nanoarchitecture.net	http://nanoarchitecture.net
최신동향	E! Science News	http://esciencenews.com
	Fresh Patents. Com	http://freshpatents.com
	Science Archived	http://www.sciencearchived.com
	Bio-Medicine	http://www.bio-medicine.org
	Biowizard	http://www.biowizard.com
	Chemeurope	http://www.chemeurope.com
	The Daily Galaxy	http://www.dailygalaxy.com
	eventseer.net	http://eventseer.net
	Atrelic Tech News	http://atrelic.com
	EDA Geek	http://edageek.com
	embedded system news.com	http://embeddedsystemnews.com
	InternetNz Blog	http://blog.internetnz.net.nz/
	Nanovip.com	http://www.nanovip.com/nanotechnology-news-hub/
	Newsletterarchive	http://newsletterarchive.org
	OriginalSignal	http://tech.originalsignal.com
	ReadWriteWeb	http://readwriteweb.com
	Science News Reviews	http://science.reviewnews.org
커뮤니티	BirdForum	http://birdforum.net
	Chemspider	http://chemspider.com
	DreamInCode.net	http://www.dreamincode.net
	Earth Portal	http://www.earthportal.org
	iMechanica	http://imechanica.org
	Spectroscopy Now	http://www.spectroscopynow.com
포럼	Bautforum	http://www.bautforum.com
	Biotechniques	http://molecularbiology.forums.biotechniques.com/forums/
	Nabble	http://www.nabble.com/
	SFN	http://www.scienceforums.net/forum/
가상 도서관	CIESIN	http://ciesin.columbia.edu
	GIS development	http://www.gisdevelopment.net
	Leonardo Energy	http://www.leonardo-energy.org/drupal/
	lightsources.org	http://lightsources.org/cms/
	Myhostas : Hosta Database	http://myhostas.net
	USGS Publication Warehouse	http://infotrek.er.usgs.gov/pubs/
	BioPortfolio	http://www.bioportfolio.com
	SOPHOS	http://cyberpulse.net

오픈코스웨어는 교육과정의 형태를 띠고 있는 무료로 공개된 온라인 디지털 출판물로 양질의 교육 자원이다(OCW Consortium). 이들은 형태상으로는 구조화된 자원으로 볼 수 있으나 접근이 용이하고 웹상의 authority를 가지고 있는 특이한 형태의 학술 자원으로 볼 수 있다. 따라서 이 연구에서는 이를 유형I 또는 유형II 자원에 분류하지 않고 별도의 자원 유형으로 보았다. OCW Consortium에서 확인한 오픈코스웨어 중 국외 15개 대학, 국내 2개 대학의 오픈코스웨어를 수집대상으로 선정하였다.

기타 사이트로는 arxiv.org의 자원을 이용하는 사이트들과 e-book을 제공하는 사이트를 포함하였다. 이 오픈코스웨어 및 기타 수집대상 사이트는 〈표 3〉과 같다.

〈표 3〉 오픈코스웨어 및 기타 후보 사이트

	사이트명	URL
자원유형	College of Eastern Utah	http://ocw.ceu.edu
	Johns Hopkins Bloomberg School of Public Health	http://ocw.jhsph.edu
	Massachusetts Institute of Technology	http://ocw.mit.edu
	Michigan State University	http://www.msuglobal.com/OpenCourseWare/
	Tufts University	http://ocw.tufts.edu
	UC Berkeley	http://webcast.berkeley.edu
오픈코스웨어	University of California, Irvine	http://ocw.uci.edu/courses/
	University of Massachusetts Boston	http://ocw.umb.edu
	University of Notre Dame	http://ocw.nd.edu
	University of Utah	http://ocw.utah.edu
	Utah State University	http://ocw.usu.edu
	Weber State University	http://ocw.weber.edu
	Open University - OpenLearn	http://www.open.ac.uk
기타	TUDelft	http://ocw.tudelft.nl
	Capilano College	http://ocw.capcollege.bc.ca
	Korea University OpenCourseWare	http://ocw.korea.edu
	Kyung Hee University	http://ocw.khu.ac.kr
	Eprintweb	http://eprintweb.org
	Science And Tech News	http://vsevcosmos.livejournal.com
	addebook.com	http://www.addebook.com

3.3 웹 페이지 수집 현황 및 KISTI 학술논문 현황

이 연구의 실험에는 학술자원으로써 KISTI의 학술논문과 위에서 선정한 학술웹사이트에서 수집한 웹페이지를 이용하였다.

실험에 사용된 자관 학술논문은 KISTI 주제분류표에 의하여 32개 주제 분야로 나누어 저장하고 있다. 총 75만7,596 건의 학술논문 중 가장 많은 것은 의약학 분야의 학술논문으로 19만257건(25.1%)이다. 그 외에는 물리 · 응용물리, 생명과학, 화학 · 화공의 주제 분야에 5만여건 이상의 학술논문이 있다. 이밖에 경영 · 경제 분야는 3만836건(4.1%), 수학 분야는 2만3,384건(3.1%), 전산 · 정보 분야는 2만993건(2.8%)의 비율로 구성되어 있다(〈표 4〉 참조).

웹 학술자원으로 수집한 웹페이지 현황은 〈표 5〉와 같다. 자체 제작한 크롤러를 이용하여 2008년 6월부터 2008년 9월에 걸쳐 37개 사이트의 총 81만9,250개 페이지를 수집하였다. 가장 많이 수집된 것은 유형II 웹 학술자원으로 28개 사이트에서 총 77만256개 페이지가 수집되었다. 유형I 웹 학술자원은 4개 사이트에서 총 1만7,932개 페이지가 수집되었으며 OCW와 기타 사이트에서는 3만1,062개 페이지가 수집되었다.

〈표 4〉 연구에 사용한 KISTI 보유 학술논문 현황

주제분류	논문수	비율(%)	주제분류	논문수	비율(%)
의약학	190,257	25.1	정치	6,650	0.9
물리.응용물리	100,254	13.2	제조	6,468	0.9
생명과학	83,891	11.1	건설.건축.토목	5,346	0.7
화학.화공	77,803	10.3	예술	4,930	0.7
식물	32,184	4.2	역사.지리	4,663	0.6
경영.경제	30,836	4.1	도시.환경	3,770	0.5
사회과학일반	28,533	3.8	천문	3,458	0.5
공학일반	27,918	3.7	법률	3,275	0.4
수학	23,384	3.1	행정	2,778	0.4
지학.지질	22,607	3	종교	2,302	0.3
동물	22,442	3	무역	2,292	0.3
전산.정보	20,993	2.8	철학	2,283	0.3
자연과학일반	14,173	1.9	언어	1,354	0.2
농업	13,890	1.8	가정.가사	1,065	0.1
교육	9,829	1.3	문학	621	0.1
심리학	6,945	0.9	기술과학일반	402	0.1
합계	757,596	100			

〈표 5〉 웹 학술자원 수집 결과

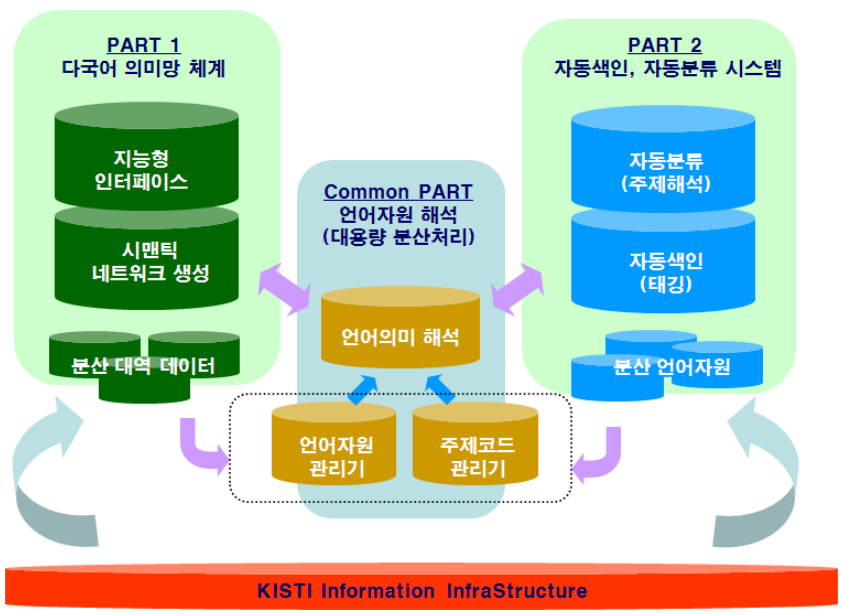
자원유형	세부유형	사이트명	수집페이지수
유형I	기관 리포지터리	Tree of Life Web Project	3,257
	온라인저널	JBC	114
		JCS	117
		WJES	14,444
		합계	17,932
유형II	블로그	The Math Mojo Chronicles	81
		Blogtica.com	3,699
		http://blogs.wsj.com/economics/	13,465
		Let's Play Math!	4,783
		LogicMatters	56,759
		Marginal Revolution	12,649
		Math-Blog	121
		Mathematics Weblog	621
		MathNotations	66,116
		MATLAB Central	745
		The Big Picture	1,776
		Wild About Math!	530
	최신동향	Best-Health-Report.com	1,541
		Bio-Medicine	52,903
		Biowizard	117,867
		ChemEurope	91,616
		eventseer.net	6,794
		The Daily Galaxy	141
	커뮤니티	BirdForum	97,780
		ChemSpider	705
		DreamInCode.net	22,082
		Earth Portal	3,496
	포럼	Bautforum	105,707
		Biotechniques	51,780
		Nabble	28,287
	virtual library	BioPortfolio	21,605
		USGS Publication Warehouse	3,062
	이용자참여 백과사전	Encyclopedia of Earth	3,545
		합계	770,256
OCW		Johns Hopkins Bloomberg School of Public Health	1,021
		Massachusetts Institute of Technology	28,928
		Michigan State University	84
		Tufts University	265
기타		addebook.com	764
		합계	31,062
합계			819,250

3.4 정보자원의 자동색인 및 자동분류 (STEAK)

STEAK 시스템은 비통제 어휘의 의미망을 이용한 다국어 질의확장 및 색인 분류를 위한 기반 시스템으로 현재 KISTI에서 개발 중인 언어자원 생성 및 분석도구이다. STEAK 시스템은 크게 두 영역으로 구분되는데 첫 번째는 다국어 어휘 간의 관련 네트워크를 자동생성하고 동적으로 해석하여 제공하는 기능이며 두 번째는 구축된 자원으로부터 언어자원의 학습 환경을 구축하고 이를 이용해 학술정보를 자동분류하는 기능이다(<그림 4> 참조). 이 연구에서는 학술논문 및 웹의 상이한 자원으로부터 주요 전문용어를 추출하고 이를 이용해 유사도를 측정하는 도구 및 시스템으로 이용하였다.

색인 및 분류를 위한 기본 언어자원을 추출하기 위해서 학습대상이 되는 KISTI 2006년 논문 약 75만 건 중 키워드 필드가 있는 약 40여만 건을 이용하였다. 키워드의 주제별 출현빈도를 바탕으로 코사인 유사계수를 이용한 주요 색인어의 주제 가중치 벡터를 생성하고 DB에 저장하였다. 또한 저장된 주제 가중치 벡터에 대해 일정한 수준의 임계치를 지정하여 분야별 전문성이 높은 데이터를 동적으로 재생산할 수 있게 하였다.

학술논문을 자동분류 하는 과정에서 유사도를 측정하기 위하여 오치아이 유사계수를 사용하였다. 오치아이 유사계수는 이진 데이터의



<그림 4> STEAK 시스템 개요도

코사인 유사계수이다(정영미 2005). 오치아이 계수를 위한 2X2 분할표는 〈표 6〉과 같다.

$$\text{오치아이 계수} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

〈표 6〉 자질(키워드)과 범주(주제 분야)간 2x2 분할표

	범주 cj 소속	범주 cj 미소속
자질 fi 출현	a	b
자질 fi 미출현	c	d

여기서 자질 f는 출현 키워드를 의미하고, 범주 c는 키워드가 속한 주제 분야를 의미한다. 문헌 자동분류 시에는 자질값(자질과 범주의 연관도) 투표방식을 사용하는데 이때, 분류대상 문서에 나타난 n개의 단어 자질집합과 후보범주 m개의 집합을 각각 F={f1, f2, …, fn}와 C={c1, c2, …, cm}로 표현한다. 자질 fi가 범주 cj에 대해서 가지는 자질값을 V(fi, cj)라고 하면 자질값 투표 분류기는 다음 공식을 만족하는 범주 cj를 문서에 할당하게 된다(이재윤 2005).

$$\arg\max_{c_j \in C} \sum_i V(f_i, c_j)$$

STEAK를 이용하여 학술논문 및 웹 정보자원을 자동태깅하고 자동분류하여 이들 간의 유사문헌을 연계하여 검색결과로 제시하였다. 유사문헌 연계방법은 KISTI에서 도입하여 운영하고 있는 검색엔진 Fast의 검색 API를 이용하여 자동추출한 색인어의 연관검색 기능을 추가 개발하였다.

3.5 검색 시스템 성능 및 전문가 검색결과 평가 방법

일반적으로 시스템 성능평가에는 정확률과 함께 시스템이 적합문헌을 검색해 내는 능력인 재현율을 함께 사용한다. 그러나 재현율은 실험용 질의에 대한 적합성 판정이 미리 내려져 있어야 하기 때문에 실제 운영 중인 시스템에서는 사용하기 매우 어렵다(정영미 2005). 따라서 이 연구에서는 검색 성능 평가에 정확률만을 사용하였다.

그리고 전문가 집단이 적합 또는 부적합을 판정하는 이용자 지향 시스템 성능 평가(김홍렬 2000)와 리커트 3점 척도를 이용하여 전문가 검색결과 평가를 실시하였다.

5개의 주제 분야에 대하여 각 주제별로 총 5명의 전문가가 평가에 참여하였으며, 질의는 전문가의 실제 정보요구에 따라 전문가가 직접 구성하여 검색하도록 하였다. 이 때 질의의 수준을 다소 일반적인 용어, 다소 전문적인 용어, 세부적인 특정 주제 분야의 용어를 모두 포함하여 총 10개 질의를 가지고 검색하도록 하였다.

논문검색 및 유사 웹 정보자원 연계 인터페이스를 개발한 후 각 주제 분야별로 전문가가 개별 아이디로 접속하여 각 분야에 해당하는 전문용어를 이용해 1차 검색을 실시한다. 1차 검색결과의 간략리스트에서 원하는 논문을 클릭하여 〈그림 5〉와 같이 상세화면으로 이동하면 논문의 초록수준의 정보와 함께 유사문헌

〈그림 5〉 논문 상세정보 및 시스템 평가 화면

으로 검색된 상위 10개의 웹문서 추천리스트가 나타난다. 상위 10개 웹문서로 평가의 범위를 정한 것은 검색인터페이스와 인간이 절대식별이 가능한 작업 기억(단기 기억)의 용량이 5~9개임을 고려한 것이다(Miller 1956). 전문가가 웹문서의 내용을 확인하고 적합성 여부를 표시하여 저장하면 1개 논문에 대한 해당 웹문서 10개에 대한 평가가 완료되도록 하였다.

4. 자동분류 및 검색결과 평가

4.1 자동분류 성능 평가

학술논문에 대해 복수 자동분류한 결과, 정확률은 〈표 7〉과 같이 약 72.7%로 나타났다. 향후 학술논문과 웹 정보자원 간 유사문헌 연계를 시도할 주제 분야인 의약학(생명과학 포함), 화학, 전산정보, 경영경제, 수학 분야의 자동분류 결과가 매우 높은 것으로 확인되었다. 이전의 실험에서 전산정보 분야는 정확률이 떨어지는 경향이 있었으나(정도헌 등 2007), 주요 학문영역이므로 본 실험에 주제영역으로 추가 선정하였다.

〈표 8〉은 자동분류에 의해 복수로 해석된 웹 정보의 주제별 분포 평균이다. 자동분류한 결과, 실험대상이 되는 주제 분야 중에서 특히 의약학(생명과학 포함), 전산정보, 경영경제의 논문 분포비율이 높게 나타나고 있다.

〈표 7〉 중복 범주 허용 시 문헌자동분류 성능

주제 분야	총건수	매칭됨	정확률(%)	주제 분야	총건수	매칭됨	정확률(%)
철학	2282	1003	43.95	동물	22441	12124	54.03
심리학	6944	2873	41.37	기술과학일반	401	138	34.41
종교	2301	902	39.20	공학일반	27917	15037	53.86
사회과학일반	28532	20658	72.40	건설.건축.토목	5345	3343	62.54
정치	6649	1991	29.94	도시.환경	3769	2286	60.65
법률	3274	676	20.65	농업	13889	8213	59.13
행정	2777	1448	52.14	가정.가사	1064	503	47.27
교육	9828	6981	71.03	제조	6467	3356	51.89
무역	2291	805	35.14	예술	4929	1214	24.63
언어	1353	637	47.08	문학	620	197	31.77
자연과학일반	14172	727	5.13	역사.지리	4662	1388	29.77
수학	23383	18776	80.30	화학.화공	77802	63784	81.98
천문	3457	2214	64.04	경영.경제	30835	26487	85.90
지학.지질	22606	14619	64.67	전산.정보	20992	11292	53.79
생명과학	83890	61787	73.65	의약학	190256	176384	92.71
식물	32183	21143	65.70	물리.응용물리	100253	68095	67.92
합계	757564	551081	72.74				

〈표 8〉 주제 분류별 웹 자원의 자동분류 결과

주제 분야	1순위 분류	2순위 분류	분포비율	주제 분야	1순위 분류	2순위 분류	분포비율
철학	20,326	14,809	2.165%	동물	14,628	8,247	1.410%
심리학	8,344	47,493	3.441%	기술과학일반	152	209	0.022%
종교	2,030	3,166	0.320%	공학일반	1,620	2,199	0.235%
사회과학일반	82,165	179,248	16.112%	건설.건축.토목	2,440	10,892	0.822%
정치	392	459	0.052%	도시.환경	295	327	0.038%
법률	607	939	0.095%	농업	161	357	0.032%
행정	5,654	14,507	1.243%	가정.가사	82	135	0.013%
교육	93,454	30,013	7.610%	제조	158	538	0.043%
무역	510	662	0.072%	예술	3,232	2,485	0.352%
언어	3,152	3,367	0.402%	문학	2,697	1,446	0.255%
자연과학일반	25,742	39,028	3.992%	역사.지리	969	1,979	0.182%
수학	14,192	47,644	3.811%	화학.화공	14,470	7,845	1.375%
천문	24,003	8,757	2.019%	경영.경제	109,594	156,176	16.380%
지학.지질	3,077	3,671	0.416%	전산.정보	130,563	68,731	12.283%
생명과학	12,099	100,199	6.921%	의약학	226,252	47,611	16.879%
식물	1,947	989	0.181%	물리.응용물리	6,272	7,109	0.825%

〈표 9〉 주제별 정확률 및 마이크로 정확률

주제 분야	측정건수	매칭됨	매칭안됨	정확률(%)
의약학	195	190	5	97.44
화학	119	58	61	48.74
경영경제	98	24	74	24.49
전산정보	175	106	69	60.57
수학	99	81	18	81.82
계	686	459	227	66.91

웹 정보자원의 자동분류를 사전에 수행하는 과정은 정보의 품질이나 내용의 편차가 매우 심한 웹 정보의 특성을 감안하여 이러한 품질 편차를 통제하기 위한 것이다. 연계대상 논문에 자동 할당된 주제 분야와 동일한 웹 정보의 주제 분야 군을 추출하여 유사 주제 분야의 군집을 만든 후 상호간 유사연계를 하는 후보 대상으로 선정하게 된다. 이와 같은 의미적인 선별과정을 통해 상호연계 과정의 품질제어와 연계성능 향상을 도모할 수 있다.

4.2 검색 성능 평가

의약학, 전산정보, 경영경제, 화학, 수학의 5개 분야 전문가에 의해 해당 논문과의 유사 문헌 적합성을 측정하였다. 각 주제별 정확도를 평균한 매크로 정확률을 측정하고, 개별질의를 모두 합하여 평균한 마이크로 정확률을 측정하였다.

〈표 9〉에서 보는 바와 같이 최종 데이터 수집결과 각 분야별로 평가수가 상이하여 분야별 건수의 차이가 발생하였다. 이로 인해 분야별 측정건수의 편차가 전체 매칭건수를 전체 건수로 나눈 마이크로 정확률에 영향을 줄 수 있으므로, 각 주제 분야별 평균정확률을 측정하여 매크로 정확률도 추가로 산출하였다.

매칭 정보자원 상위 10개의 전체 정확률인 마이크로 정확률은 66.91%로 나타났다. 그리고 매칭 정보자원 상위 10개의 주제별 정확률의 평균인 매크로 정확률은 62.61%를 보여, 두 가지 측정결과가 크게 상이하지 않았다.

논문의 자동분류 결과에서 매우 우수한 결과를 보였던 주제 분야인 의약학(92.71%), 수학(80.30%)은 매우 성능이 좋은 것으로 나타났으나, 논문 자동분류에 비해 성능이 저조한 분야는 화학(81.98%)과 경영경제(85.90%)로 특히 경영경제가 저조한 것으로 나타났다. 향후 경영경제 분야의 용어에 대한 해석기법을 재검토하고, 경영경제 분야의 웹 자원을 충분히 증가시켜 재실험을 해야 할 것으로 보인다. 또한, 일반적으로 응용분야의 성격이 강해 자동분류 성능이 저조하게 나타나는 전산 · 정보 분야는 논문 분류결과(53.79%)와 유사한 수준인 60%대에 머물렀다.

4.3 전문가 검색결과 평가 결과

5개 전문 분야별로 시스템이 제시한 항목 리스트마다 3점 척도를 부여하고 이를 합산하는 방식으로 전문가 검색결과 평가를 측정하였다.

전문가 검색결과 평가를 전체 건수를 기준으로 측정한 결과 〈표 10〉과 같이 약 76.8%로 비교적 우수한 것으로 나타났다. 또한, 분야별로 측정한 결과는 의약학, 수학분야가 높고 전산정보, 화학, 경영경제 순으로 만족하는 것으로 나타났다. 이는 앞서 수행한 정확률 측정결과에서 보이는 순위(의약학〉 수학〉 전산정보〉 화학〉 경영경제 순)와 일치하는 결과를 보이고 있다는 점에서 중요하다고 할 수 있다. 즉, 실제 전문가 검색결과 평가는 의미적으로 매칭을 잘 수행하는 경우에 높게 평가되었다는 것이다. 따라서 서비스 만족도를 높이기 위해서는, 다른 유형의 정보자원 간 매칭결과의 품질을 향상시켜 시스템 해석성능을 높이는 것이 가장 중요하다고 할 수 있다.

5. 결론

이 연구는 웹 학술자원을 수집하여 자동으로 연계하는 시도를 통해 기존의 학술논문 서비스에서 더욱 확장된 학술정보서비스를 제공하기 위한 서비스 기술을 개발하기 위한 것이다. 이를 위해 의미색인을 사용하여 언어자원의 특성이 이질적인 자원들 간의 연계를 시도하여 의미연계 서비스의 가능성을 제시하고자 하였다. 이 연구의 시도는 다양한 웹 학술자원의 이용가능성에 대해 점검해 볼 수 있는 계기가 되었으며 지속적인 웹 자원의 학술적 활용에 대한 가능성을 제시했다고 할 수 있다. 이 연구의 결과는 다음과 같다.

검색 성능은 매크로 정확률과 마이크로 정확률 모두 60% 이상의 성능을 보였다. 또 NDSL 제공 학술논문을 자동분류한 결과가 우수했던 의약학과 수학 주제 분야의 성능이 80% 이상으로 매우 좋게 나타났다. 이를 통해 볼 때 STEAK 시스템의 유용성을 확인할 수

〈표 10〉 전문가 검색결과 평가

	상	중	하	합계	전체
의약학	188	7	0	195	98.8%
화학	10	79	30	119	61.1%
경영경제	0	48	50	98	49.7%
전산정보	68	59	48	175	70.5%
수학	70	29	0	99	90.2%
합계	336	222	128	686	76.78%
백분율(%)	49.0%	32.4%	18.7%	100%	-

있었으며, 후속 연구를 통해 성능이 미비한 분야의 웹 자원을 보강하는 것이 필요할 것이다.

전문가 검색결과 평가는 70점 이상으로 비교적 우수하게 나타났으며, 시스템 성능과 마찬가지로 의약학 및 수학 분야에서 높은 전문가 검색 평가를 보이고 있었다. 실제 전문가 검색결과 평가는 시스템 성능이 높은 경우 즉, 의미적으로 매칭을 잘 수행하는 경우에 높게 평가되었다. 이를 통해 볼 때 서비스 만족도를 높이기 위해서는 다양한 형태의 정보자원 간 매칭 결과를 향상시켜 시스템 해석 성능을 높이는 것이 가장 필요한 것으로 보인다.

이 연구에서 구축한 시스템은 다음과 같은 점에서 기존의 연구들과 차이점을 가진다.

첫째, 웹 학술 자원을 선정함에 있어서 기존의 연구들이 전자저널 및 연구자 목록, 타 학술정보기관 사이트 등을 주로 고려한 것에 비하여, 이 연구에서는 블로그 및 커뮤니티 사이트, 웹 포럼사이트와 같이 비교적 다양한 정보원을 이용하고자 하였다. 이는 웹 학술자원의 범위가 기존의 심층 웹(hidden web, invisible web, deep web)(Bergman 2001; D. Florescu et al. 1998)에 대한 관심을 넘어서서 현재 발현하고 있는 다양한 표층 웹(visible web)상의 유용한 정보에의 접근성 및 권위(authority)를 고려해보고자 하는 시도로 해석할 수 있다. 또한 이는 심층 웹의 기존 정의에 블로그 및 소셜 네트워킹 서비스 사이트를 추가하는 시도로도 볼 수 있다.

둘째, 의미색인 시스템인 STEAK를 도입함으로써 웹 자원에서 발생할 수 있는 노이즈를 줄이고 KISTI 자원과 의미적으로 유사한 웹 학술정보원을 검색결과로 제시하고자 하였다. 또한 STEAK의 의미해석시스템을 이용하여 학습을 수행함으로써 웹 학술자원이 지속적으로 추가되더라도 자동 학습을 통해 유의미한 색인어를 추출할 수 있으므로 이를 통해 시스템적으로 NDSL 학술자원과의 의미색인어 상호해석이 가능할 것이다.

셋째, 검색결과를 제공할 때 일차적으로 NDSL 학술자원을 보여주고 이용자가 선택한 NDSL 학술자원과 유사한 웹 학술자원을 보여주는 방식을 취함으로써 검색결과 통합에서 발생할 수 있는 오류를 최소화하고자 하였다.

이 연구의 유용성 및 후속 연구 가능성은 다음과 같다.

첫째, 웹 자원을 학술적으로 이용하고자 하는 시도 중에서도 기계 학습을 통해 의미색인을 시도함으로써 학술논문 정보원과 웹 학술정보원을 모두 사용하는 효율적인 시스템을 제안하고자 하였으며 이는 KISTI의 정보유통 서비스에 있어서 유용하게 활용될 수 있을 것이다.

둘째, 블로그 이용률 현황을 고려할 때 KISTI 서비스의 잠재적 이용자층으로 볼 수 있는 청소년들의 이용을 더욱 활성화할 수 있는 서비스 제안이 가능하다.

향후 추후 웹 정보자원의 유형 및 수집 사이트를 확대하고 주제별로 보완하여 의미색인에 의한 자동분류 결과를 자세히 분석하고 검

색과 연결시키는 후속연구가 가능할 것으로 기대한다.

참고문헌

김홍렬. 2000. 적합성 평가기준 변화에 관한 실험 연구. 『한국도서관 · 정보학회지』, 31(4): 139-164.

윤정옥. 2003. 웹 자원의 서지적 조직과 접근: 국내외 대학 도서관의 사례연구. 『정보관리학회지』, 20(1): 271-299.

이재윤. 2005. 문서측 자질선정을 이용한 고속 문서분류기의 성능향상에 관한 연구. 『정보관리연구』, 36(4): 51-69.

정도헌. 2007. 다국어 전문용어의 의미망을 이용한 질의확장 시스템 구현. 『제10회 디지털 도서관 컨퍼런스』. 2007년 11월 29일. [서울: 코엑스 그랜드볼룸].

정도헌, 김환민, 김혜선, 신기정. 2007. 과학기술 전문용어의 주제 분야별 전문성과 자동분류 성공률 간의 연관성 비교. 『제14회 한국정보관리학회 학술대회 논문집』, 31-36.

정도헌, 최희윤. 2006. 과학기술 전문용어의 다국어 의미망 생성과 분석. 『정보관리연구』, 37(4): 25-47.

정영미. 2005. 『정보검색연구』. 서울: 구미무역.

Baum, J.A.C., Calabrese, T., and Brian S. Silverman. 2000. "Don'T Go It Alone: Alliance Network Composition and Startups' Performance in Canadian Biotechnology." *Strategic Management Journal*, 21(3): 267-294.

Bergman, Michael K. 2001. The Deep Web: Surfacing Hidden Value, Deep Content White paper. [cited 2008. 09. 18]. 〈http://beta.brightplanet.com/deepcontent/tutorials/DeepWeb/index.asp〉.

Bloehdorn S. and Andreas Hotho. 2006. "Boosting for Text Classification with Semantic Features." *Lecture Notes in Computer Science*, 3932: 149-166.

Florescu, D., Levy, A., and A. Mendelzon. 1998. "Database Techniques for the World-Wide Web: A Survey." *SIGMOD Record*, 27(3): 59-74.

Koch, T. 2000. "Quality-Controlled Subject Gateways: Definitions, Typologies, Empirical Overview." *Online Information Review*, 24(1): 24-34.

Liebeskind, J. P., Oliver, A. L., Zucker, L., and Marilynn Brewer. 1996. "Social Networks, Learning, and Flexibility: Sourcing Scientific Knowledge in New Biotechnology Firms." *Organization Science*, 7(4): 428-443.

Miller, G. A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review*,

63: 81-97.

Moschitti, A., and Roberto Basili. 2004. "Complex Linguistic Features for Text Classification: A Comprehensive Study." *LNCS*, 2997: 181-196.

Owen-Smith, J., Riccaboni, M., Pammolli, F., and Walter W. Powell. 2002. "A Comparison of U.S. and European University-Industry Relations in the Life Sciences." *Management Science*, 48(1): 24-43.

OCW Consortium. 〈http://www.ocwconsortium.org/〉.