시소러스와 분야분류체계를 이용한 과학기술문헌에의 주제 및 분야할당

정한민° 강인수 성원경
한국과학기술정보연구원 NTIS사업단
{jhm°, dbaisk, wksung}@kisti.re.kr

Assigning Topics and Categories to Science & Technology Documents Using Thesaurus and Category Map

Hanmin Jung°, In-Su Kang, Won-Kyung Sung
NTIS Division, KISTI

요 약

본 연구는 문헌으로부터 추출한 색인어를 시소러스 개념어와 매칭시킴으로써 해당 문헌을 대표하는 주제 및 분야할당이 가능하도록 하는 자동 주제 및 분야할당 기법을 제안한다. 기존에는 특정 주제나 분야를 대표할 수 있는 용어 집합으로 구성된 사전을 이용하는 원시적 방법이나 문헌 내의 색인어들 중 중심이 되는 것이라고 판단되는 색인어를 통제 없이 주제로서 제시하는 연구들이 이루어졌다. 한편 기업체를 중심으로 한 KMS 시스템들은 이러한 자동화된 기법을 배제하고 수작업으로만 문헌들에 주제나 분야를 직접적으로 할당하고 있으나 이 역시 비용 측면이나 유지보수 측면에서 정보확장에 그 한계가 있을 수 밖에 없다. 빠르게 증가하는 문헌들을 효율적으로 분류하고 서비스하기 위해서는 잘 통제된 주제집합을 이용할 필요가 있으며, 주제뿐만 아니라 분야 관점에서의 접근 또한 필요하다. 본 연구는 이러한 요구사항들을 만족시킬 수 있도록 시소러스라는 통제된 어휘집합을 중심에 두고 정보검색시스템에서의 색인어와 분야분류체계의 분야분류명들을 상호 연계하는 방식으로 주제 및 분야할당 시스템을 제안한다. 시소러스 개념어들과 분야분류체계의 분야분류명들간에 매핑이 이루어지도록 언어자원을 구성하며, 색인어와 개념어 매칭 결과를 활용하여 주제 및 분야를 해당 문헌에 자동 할당할 수 있도록 한다. 과학기술문헌에 대한 주제 및 분야할당을 위해 15,000 여 개의 범용 과학기술 시소러스 개념어들을 사용하고 있으며, KOSEF 분야분류체계를 이용하여 시소러스와의 매핑을 시도하였다. 본 연구를 통해 구현된 주제 및 분야할당 시스템을 이용하여 문헌에 할당한 주제 및 분야정보는 추론규칙을 이용하여 온톨로지 기반의 지식으로 변환되고 추론 서비스를 통해 연구자에게 제공된다.

KEYWORDS: 시소러스, 분야분류체계, 주제, 분야, 색인, 시맨틱 웹, 정보유통, OntoFrame-K®

1. 서론

최근 정보검색 분야에서 자연어처리 기술은 형태소분석 및 구문분석을 통하여 품사의 식별과 원형 그리고 문장 내에서의 역할 등을 상당 수준 검출할 수 있게 발전하였다. 그러나 정보통신 및 정보기술의 발달로 새로운 분야 및 용어가 급격히 증가하는 추세에서 종래의 정보검색 방법을 적용하는 경우, 새로운 주제나 분야에 따라 잘 분류된 정보를 제공하는 것이 어렵다. 이는 기존 정보검색에서는 대부분 수작업에 의해 문헌에 주제 및 분야를 직접 할당하거나, 통제되지 않은 어휘를 자동방식을 통해 주제어로서 제시하는 등의 체계적이지 않은 방법들을 사용하여 지속적이고 일관성 있는 정보획득 및 고품질의 서비스를 어렵게 만들고 있기 때문이다. 즉, 종래의 정보검색 방법에서는 정보의 증가 속도에 따라 시스템 관리자 또는 정보 관리자가 수시로 문헌 내 주제 및 분야를 변경하고 할당해야 하는데, 이는 수많은 비용과 인력을 필요로 한다. 더욱이 관리자의 컨디션이나 업무인수·인계 등의 환경적 요소도 중요하게 작용하여 일관성 없는 주제 및 분야할당이 이루어질 수 있다는 심각한 문제도 내포하고 있다. 한편으로는 색인어로부터 자동으로 주제를 선정하는 시도들이 있었지만, 이

역시 색인어 통제가 이루어지지 않거나 단순한 Bag of Words 개념을 사용함으로써 부적절한 용어를 주제어로 제시할 수 있다는 문제점을 가진다.

따라서 본 연구는 상기 문제점들을 해결하고자 자동 주제 및 분야 결정 기법과 통제된 개념어로 구성된 시소러스를 활용하여 효율적이면서도 주제어로서 부족하지 않은 개념어들을 문헌에 자동 할당할 수 있도록 한다. 분야할당을 위해서는 색인어와 매칭이 용이한 개념어를 매개체로 하여 개념어에 분야분류명들을 매핑하여 간접적으로 색인어와의 매칭을 시도한다. 본 주제 및 분야할당 시스템은 시소러스나 분야분류체계가 변경되어야 하는 경우에 해당 언어자원의 갱신 후 정보검색시스템에서 기본적으로 제공하는 전체색인을 수행하여 자동적으로 문헌들에 할당된 주제 및 분야를 일관성 있게 변경할 수 있다는 장점을 가진다. 또한 본 연구는 정보검색시스템을 이용하여 관련 문헌뿐만 아니라 관련 분야 및 주제에 대한 다양한 정보들을 함께 제공함으로써 고품질의 정보검색 및 정보분류 결과를 연구자에게 서비스하는 효과를 얻을 수 있다.

2. 기존연구

(안찬민 외 2004)는 이메일 분류를 위해 색인어에 기반한 동적분류체계를 이용하였다. 이 방식은 메일이 분류될 메시지의 주제가 자동 생성됨으로 사용자의 간섭이 필요 없고 동적분류체계를 사용하여 유연한 이메일 재분류 및 디렉터리 검색이 가능하다는 장점을 가지는 반면에, 주제어가 색인어에서 자동으로 선정됨으로 인해 부적절한 주제어를 포함할 가능성이 커지고[1] 한정된 주제로서 문헌이나 이메일들을 통제할 방법을 제공할 수 없다는 문제점을 가진다. (정호석 외 2000)은 문서 자동분류 결과로서 제시하는 분류 주제어를 수작업이 아닌 자동으로 획득하는 방식을 제안하였다. 분류 주제어를 수작업으로 구축하는 시간과 비용을 절감하고자 후보 주제어들에 대해 응집도를 계산하고 평가하여 새로운 주제어로서 추가하는 자동화된 공정을 보여준다. 그렇지만, 이들이 문제로 제기한 분류 주제어는 단순한 Bag of Words 수준을 생각한 것으로 시소러스는 이러한 주제어 제시 이외에도 확장검색, 용어 정형화 등 다양한 분야에서 사용되는 핵심적인 언어자원이다. 또한, 분야 관점에서의 문헌 분류가 이루어져야 할 경우에 이 시스템은 별도의 신규 시스템을 설계하고 구현해야 하는 이중적인 관리가 되어야 한다. 이러한 별개 시스템은 유지보수를 오히려 더 어렵게 만들고 서로간의 연관성을 배제함으로써 문헌 분류에 있어서의 시너지 효과를 얻지 못한다는 약점을 가질 수 밖에 없다.

체계적인 시스템 통제를 위해 시소러스를 이용하는 연구들이 여러 응용분야에서 수행되었다. (이창범, 박혁로 2001)는 시소러스 상의 관계 (USE/UF 관계, 동등관계 등)들을 이용하여 문헌 내에서 출현한 개념어들에 차별적으로 가중치를 부여하여 중요 문장을 발견하고 이를 요약문으로 제시하는 기법을 제안하였다. 그렇지만, 이 기법은 주제 제시에는 간접적으로 활용할 수 있지만 분야 제시를 위해서는 사용할 수 없으며, 주제나 분야로서 부적절한 개념어들에 대한 불용어처리가 이루어지지 않고 있다. 시소러스를 이용한 또 다른 연구로는 (방선이, 양재동, 양형정 2004)가 있다. 여기에서는 범주 (주제)를 결정하기 위한 세부 범주사전을 구축하고 k-NN 분류 알고리즘을 이용하여 문헌들을 자동분류하고자 하였다. 특히 특정 범주로 분류하기가 어려운 경우에 시소러스 상의 각 관계들에 부여된 가중치를 이용하여 분류의 정확도를

[1] 4.4절에서 보듯이 통제된 어휘집합인 시소러스를 사용하는 경우에도 분야로서 부적절한 개념어들을 약 43.5% 포함하고 있다.

향상시키고자 하였다. 그러나 세부 범주사전은 구축자의 상식, 관점 등에 크게 의존하는 주관적 정보를 담고 있으며, 이를 지속적으로 유지보수하고 확장하는 데 있어서의 일관성을 획득할 근거가 없다. 이 연구 역시 분야 관점에서의 정보제공은 고려하지 않고 있다. 시소러스를 이용하지는 않지만 상기 연구에서 사용한 세부 범주사전과 유사한 방식의 장르 및 장르 내 주제범주 테이블을 사용한 연구로 (이용배, 맹성현 2003)이 있다. 장르 관점의 문헌분류를 제공하고자 문서 형식과 스타일정보를 활용하고 있다는 특징을 가진다. 그렇지만, 상기 연구와 마찬가지로 체계적이지 못한 장르 내에서의 주제 별 범주정보를 이용함으로써 지속적이며 일관성 있는 정보확장에는 제약이 따를 수밖에 없다.

3. OntoFrame-K®: 시맨틱 웹 기반 정보유통 플랫폼

지식 기반 정보유통 플랫폼 (OntoFrame-K®)은 정보 공유 및 유통 기술과 시멘틱 웹 기술이 융합된 새로운 개념의 정보유통 플랫폼으로서, 이미 검증된 전문성과 높은 부가가치를 지닌 과학기술 지식정보를 공유하고 유통시킬 수 있는 자발적 가상 협업 연구 커뮤니티 (Self-organizing Virtual Research Community)의 구현을 지원하기 위한 시스템이다. 시멘틱 웹 기술은 유통 대상 정보인 지식정보의 모델링에서뿐만 아니라 다양한 추론 기반 응용 서비스들에도 적용되어 데이터들 간의 연관 관계를 보다 효율적으로 제시함으로써 협업 연구 커뮤니티를 통한 정보획득 과정의 효율화에 활용되고 있다. 특히 시멘틱 웹 기반 서비스의 핵심 기반 지식자원으로 활용되는 온톨로지와 시소러스는 각각 서비스 종속적 지식과 언어 종속적 지식의 모델링과 추론 과정에 활용되고 있으며, 전체 과학기술 분야를 대상으로 한 국가 과학기술 R&D 기반정보 온톨로지 (National Science & Technology R&D Reference Ontology)와 범용 과학기술 분야 시소러스가 구축되었다.

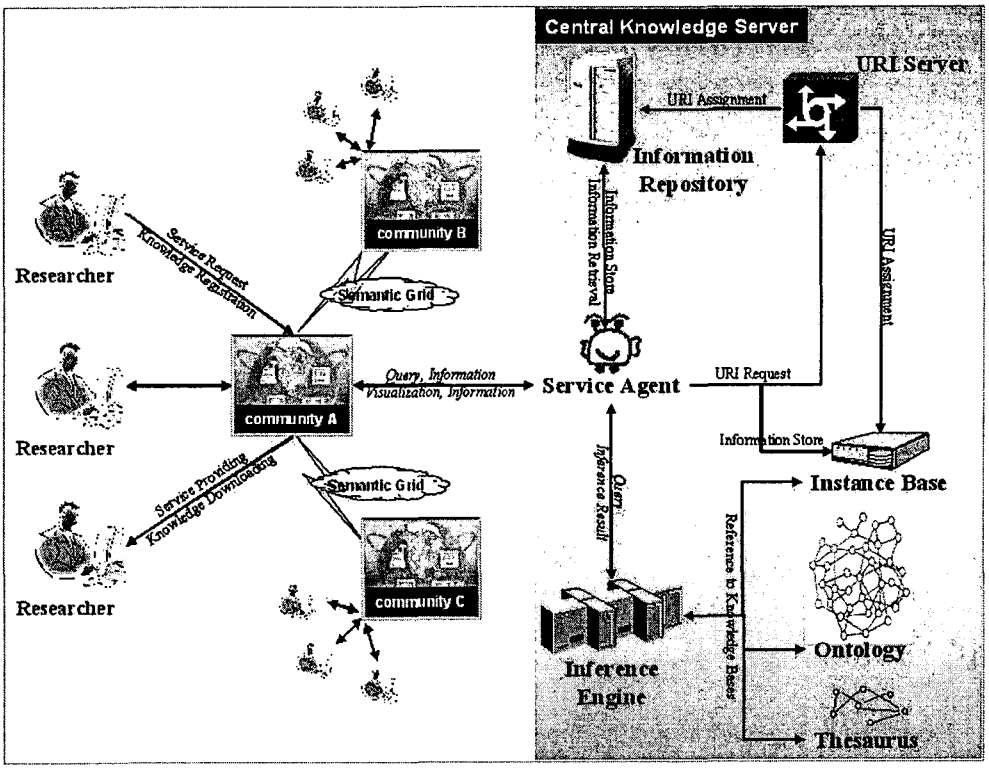

그림 1. 정보유통 플랫폼 상에서의 정보유통 서비스와 추론 서비스의 연관도

OntoFrame-K®에서 제공하는 서비스는 정보유통 서비스와 추론 서비스로 나누어진다. 연구자들 간의 자발적인 협업 지원을 위한 정보유통 플랫폼으로서의 OntoFrame-K®는 중앙 지식서버를 통해 신뢰성 있는 정보만을 교환할 수 있도록 클라이언트-서버 모델로 구성된다 (그림 1). 각 연구자인 클라이언트는 중앙 지식서버를 통해 지식정보를 등록하기도 하고 다른 연구자가 등록한 지식정보를 내려 받을 수도 있다. 중앙 지식서버는 등록된 지식정보를 검증하고 해당 정보의 이력을 추적하며, 온톨로지 기반의 추론을 통해 연구자들에게 국가 과학기술 R&D 기반정보에 기반한 추론 서비스를 제공한다.

4. 주제 및 분야할당

문헌에 수작업으로 주제 및 분야를 할당하는 방식은 주제어 목록이나 분야분류체계가 변경되는 경우에 유지보수를 어렵게 하고, 지속적인 정보확장에 있어서도 인적·물적자원에 대한 상당한 제약을 가져온다. 콘텐츠와 요소를 분리하고자 하는 시도가 온톨로지를 기반으로 하는 시맨틱 웹의 등장과 함께 본격화되고 있으며, 이러한 맥락에서 문헌에의 주제 및 분야할당 역시 새로운 방식으로 접근할 필요가 있다. 이에 본 연구에서는 주제어로 사용될 수 있는 시소러스 개념어에 분야분류명을 매핑시킨 형태의 통합적 언어자원을 이용하여 자동적으로 문헌에 주제 및 분야를 할당하는 방안을 제시한다. 본 장에서는 주제 및 분야할당을 위한 기초 언어자원인 시소러스와 분야분류체계를 시작으로 이들을 이용하여 어떠한 방식으로 주제 및 분야할당이 가능할 수 있는 지를 설명하는 알고리즘과 그 예까지 순차적으로 기술한다.

4.1 시소러스 구축

특정 전문분야에서 사용되는 빈도수가 높은 통제된 어휘집합으로 정의되는 시소러스는 정보의 급증과 검색환경의 변화에 따른 기능변화가 절실히 요구되는 자원이다. 즉, 정보의 색인 및 확장검색 (Query Expansion), 상세검색 (advanced search) 또는 추론검색 (inference search) 등의 검색 시스템에서 높은 효율성과 정확성을 보장할 수 있어야 한다. 시소러스는 전통적으로 문헌정보학과 전산학이 주축이 되어 구축되어 왔고, 전자가 여러 용어 간 관계기술 및 구축에 명확한 의미적 준거를 고려하지 않았다면, 후자는 주로 통계적 기법에 입각한 자동구축에 비중을 두어왔다. 그렇지만, 시소러스 구축 자체는 수동구축과 자동구축 모두에 있어 여러 어려운 점들을 가지고 있다. 이를 요약하면 다음과 같다.

(1) 시소러스를 수동구축할 경우 용어의 구축은 물론이고, 지식 및 정보의 증가 및 변화로 인한 향후 유지보수에 시간과 전문가의 수작업이 과도하게 요구된다. 또한 자동구축 또는 기 구축된 어휘의미망을 이용한 간접 구축의 경우라도 뚜렷한 정제 기준을 바탕으로, 지속적인 수정과 정제가 불가피하다.

(2) 시소러스가 담고 있는 지식은 지속적으로 변화, 발전하므로 시소러스가 구축됨과 동시에 시소러스에 포함된 지식은 동시대의 지식이 아닌 낡은 지식이 반영될 우려가 있다. 따라서, 한번 구축한 시소러스의 유지와 지속적인 수정이 가능하도록 초기부터 지능적으로 고안, 설계되어야 한다.

(3) 자동 구축 시 데이터의 산발성 (Data Sparseness)이 문제가 된다. 통계기법에 의한 구축관련 선행연구들에 따르면 의미적 유사성을 지닌 용어들은 유사한 통사적 관계 속에 출현한다는 가정에서 출발하며, 용어 각각은 출현 가능한 문법적 문맥에 의해 그룹화된다. 그런데, 시소러스의 구축 대상이 되는 전문용어는 코퍼스 상에서 상대적으로 낮은 빈도를 보이기 때문에 코퍼스를 이용하여 통계적으로 의미 있는 정보를 추출하기란 사실상 어렵다. 이점은 자동 구축에 의한 시소러스가 실제로 활용되는 예가 전무하다는 사실과 자동 구축 자체가 실험적 수준을 벗어나기 어렵다는 한계를 반증하는 부분이다.

(4) 향후 구축되는 시소러스는 기 구축 용어는 물론이고, 이를 기반으로 확장된 용어 및 외래어로부터 음차된 용어, 새로 생성된 신조어 (Neologism) 등을 충분히 반영할 수 있도록 유연성 있는 설계가 요구된다. 또한, 추론 서비스 등으로의 상호운용성 염두에 두어야 한다.

본 연구는 시소러스의 자동구축이 갖는 문제점 해결을 위한 대안으로, 의미적 준거를 이용한 직접 구축을 채택한다. 이를 위해 개념패싯과 관계패싯을 정의하여 개념어추가에 직접적으로 활용한다. 개념패싯은 개념어가 갖는 대표적인 의미속성, 즉 범주를 의미한다. 본 연구에서는 15개 개념패싯을 선정하였다[2]. 개념패싯 할당은 모든 개념어에 대해 이루어지며 개념 패싯이 상이한 동형이의어에 대해서는 개념분화를 실시한다. 관계패싯은 용어 간 또는 상·하위어 간의 의미관계를 표현하는 메타 개념으로서 속성관계패싯, 범주관계패싯, 의미역관계패싯, 속성키워드 등으로 나누어 계층 구조에 반영한다 (황순희, 정한민, 성원경 2005) (Jung, Sung, and Park 2006).

2005년 범용 과학기술 시소러스로서 15,000여 개 개념어에 대한 구축을 완료하였으며, 2006년 현재 15,000여 개 IT 분야 용어들을 대상으로 시소러스를 구축하고 있다. 본 연구에서는 2005년에 기 구축된 범용 과학기술 시소러스를 대상으로 주제 및 분야할당 실험을 수행하였다.

4.2 분야분류체계 선정

범용 과학기술 시소러스 상의 개념어들과 분야분류체계를 매핑하기 위해서는 먼저 과학기술 분야분류체계가 필요하다. 현재 우리나라에서 널리 사용되고 있는 대표적인 5개 분야분류체계들을 비교하면 표1과 같다.

---

[2] 본 연구에서 설정한 15 개 개념패싯은 연구의 관점에 따라 연구자들 사이에 쉽게 합의를 보기 어려운 부분이다. 이를 위해 본 연구는 다음의 절차를 거쳐 개념패싯을 설정하였다. 먼저 기존의 문헌정보학에서 설정한 20 여 개 개념패싯을 전체 용어 중 15% 내외 용어에 할당해 보았다. 할당 결과, 설정된 개념패싯의 잉여성 또는 불충분성 등을 고려하여 개념패싯을 재조정하였고, 이와 동시에, 1 차 선정된 개념패싯의 타당성 여부를 검증하기 위해, 어휘의미망의 대표적 사례인 워드넷 (Princeton WordNet)에 사용된 상위 개념(Upper Concepts)과 SUMO 의 상위 온톨로지를 참고하였다. 즉, 상위 개념을 따로 추출하여 이들로 해당 용어의 기술이 가능한지 검토하고, 1 차 선정된 개념패싯과 이를 비교하였다. 마지막으로, 용어

표 1. 과학기술 분야분류체계 간 비교

구분	한국과학재단 (KOSEF)	한국과학기술 기획평가단 (KISTEP)	한국학술진흥재단 (KRF)	정보통신 연구진흥회 (IITA)	한국산업 기술평가원 (ITEP)
범위	■ 자연과학, 공학, 생명과학 ■ 인문사회, 예체능 미포함	■ 자연과학, 공학, 생명과학 ■ 인문사회, 예체능 미포함	■ 예체능까지 포함하는 모든 분야	■ 정보통신 관련분야	■ 산업기술 관련분야
특징	■ 연구활동중심 ■ 기술공급중심	■ 과학기술중심 ■ 신생 및 융합기술반영	■ 연구활동중심 ■ 기술공급중심	■ 정보통신기술중심	■ 산업기술중심
분류 체계	■ 대분류: 10 ■ 중분류: 131 ■ 소분류: 941	■ 대분류: 19 ■ 중분류: 177 ■ 소분류: 1,229	■ 대분류: 8 ■ 중분류: 152 ■ 소분류: 1,566 ■ 세분류: 2,506	■ 대분류: 9 ■ 중분류: 31	■ 대분류: 5 ■ 중분류: 45 ■ 소분류: 414
단점	■ 인문사회 분야분류체계 없음	■ 일반연구관리에 부적합	■ 세분류 중복이 많음 ■ 분야간 연계 폐쇄적	■ 특정분야분류체계	■ 특정분야분류체계

표 1에서 보는 바와 같이 KRF의 분야분류체계는 너무 많은 분야분류를 가지고 있으며 과학기술뿐만 아니라 예체능까지 포함되어 있고, IITA와 ITEP의 분야분류체계는 정보통신 관련분야와 산업기술 관련분야라는 특정분야에 집중한 분야분류체계이므로 범용 과학기술에 부합하지 않아 배제하였다.

선정대상을 일단 KOSEF와 KISTEP을 압축하고 매핑 적합성과 활용도 분석을 위해 선행매핑을 수행하였다. 이를 위해 시소러스 15개의 개념패싯들 중 과학기술 분야 개념어들을 가장 많이 포함하고 있는 '내용' 개념패싯을 대상으로 하여, Depth 2에 속한 153개의 개념어들을 이용하였다. 표 2는 선행매핑 결과를 보여준다.

매핑 유형	KOSEF 분야분류체계	KISTEP 분야분류체계
특정 과학기술 분야	80 (52.3%)	66 (43.1%)
범용 과학기술 분야	30 (19.6%)	44 (28.8%)
일반 분야	43 (28.1%)	43 (28.1%)
총합	153	153

DB 자체의 실증적 분석 및 워드넷 상위 개념을 이용하여 적은 수의 개념패싯으로 많은 수의 용어를 효율적으로 기술할 수 있도록 개념 유형을 조정하였으며, 결국 15 개 개념패싯을 선정하였다.

표 2에서 알 수 있듯이, KOSEF 분야분류체계와 정확히 매핑되는 개념어의 수가 더 많다. 반면 KISTEP 분야분류체계에 대한 매핑 결과가 상대적으로 떨어지는 이유는 과학기술 분야뿐만 아니라 산업기술 분야를 많이 포함하고 있어서 분야분류체계에 대한 이해조차 어려운 경우가 많았으며 많은 분야분류가 너무 구체적이거나 또는 너무 포괄적이어서 매핑에 어려움이 많았기 때문이다. KOSEF 분야분류체계는 학문 분야분류체계에 속하기 때문에 상대적으로 분야분류체계를 이해하는 것이 더 용이하여 본 연구를 위한 언어자원으로 사용하기에 적합하여 이를 매핑대상 분야분류체계로 선정하였다.

4.3 시소러스-분야분류체계 간 매핑

시소러스-분야분류체계 간 매핑은 분야분류체계에 속한 분야분류명들을 시소러스 개념어들에 매핑한다는 것을 의미한다. 특정 문헌의 주제를 선정하기 위해 시소러스 개념어를 주제어로서 사용하는데, 분야 선정을 위해서는 분야분류체계가 필요하다. 그렇지만, 자동화 과정을 거쳐 문헌에 동적으로 주제 및 분야를 할당하기 위해서는 문헌 내 정보를 이용할 필요가 있다. 문헌은 결국 텍스트로 구성되어 있고, 텍스트를 구성하는 것은 키워드 (색인어) 집합이므로 키워드와 매칭 가능한 시소러스 개념어를 이용하는 것이 효율적이다. 그렇지만, 시소러스 개념어는 주제를 대표하는 주제어로서 활용가능 하지만 분야분류체계의 분야분류명들은 매칭 대상을 문헌으로부터 찾을 방법이 없다. 이에 시소러스 개념어를 매개체로 하여 문헌과 분야분류명을 매칭한다면 문헌에 주제뿐만 아니라 분야를 할당하는 것이 가능해진다[3]. 본 연구는 이러한 분야할당 방법을 도입하고자 먼저 시소러스 개념어에 분야분류명들을 할당하는 매핑을 수행한다. 하나의 시소러스 개념어에 반드시 하나의 분야분류명만 할당되는 것이 아니라 0 ~ n[4]개의 분야분류명이 할당될 수 있다. 정교한 매핑이 이루어질 수 있도록 시소러스 개념어들을 다음의 7가지 매핑 유형으로 나눈다.

(1) 소분류 단일 일치어: 다른 소분류에서는 사용되지 않고 오직 하나의 소분류에서만 사용되는 개념어 (예. "자바개발툴")
(2) 소분류 다중 일치어: 한정된 몇 개의 소분류에서만 사용되는 개념어 (예. "나노컴퓨터시스템")
(3) 소분류 공통어: 특정 중분류 내의 모든 소분류들에서 공통적으로 사용되는 개념어 (예. "전자")
(4) 소분류 판단 보류어: 중분류 구분은 가능하나 해당 중분류 내에서 특정 소분류를 선택하기 어려운 개념어 (예. "이어폰")
(5) 소분류 추가 필요어: 특정 중분류 내에서 새로운 소분류로서 추가가 필요한 개념어 (예. "로봇")
(6) 개념파악 불능어: 의미파악이 되지 않는 개념어
(7) 매핑대상 제외어: 전체 분류에 공통적으로 사용되는 개념어 또는 어느 분류에도 매핑되지 않는 개념어 (예. "시스템")

---

[3] 특허 출원 번호: 10-2006-0014749 (출원일: 2006.2.15) - 시소러스 매칭에 의한 문서DB 형성방법 및 정보검색 방법
[4] 분야와 무관한 용어의 경우 0개의 분야분류명이 할당될 수 있으며, 모든 분야에 해당하는 경우에는 전체분야분류명 개수만큼 할당될 수 있다.

그림 2는 이러한 매핑유형이 어떻게 결정될 수 있는 지를 보여주는 Workflow이다. 현재 하나의 개념어에 최대 3개까지의 분야분류명을 할당할 수 있도록 하였다. 또한, 최우선 소분류는 신중히 결정하되 나머지 할당된 소분류들은 우선순위를 두지 않도록 하였다. 이는 응용분야에 따라 하나의 소분류만을 사용할 수도 있다는 가정에 따라 해당 개념어에 가장 적합한 소분류를 1순위로 할당하고자 함이다.

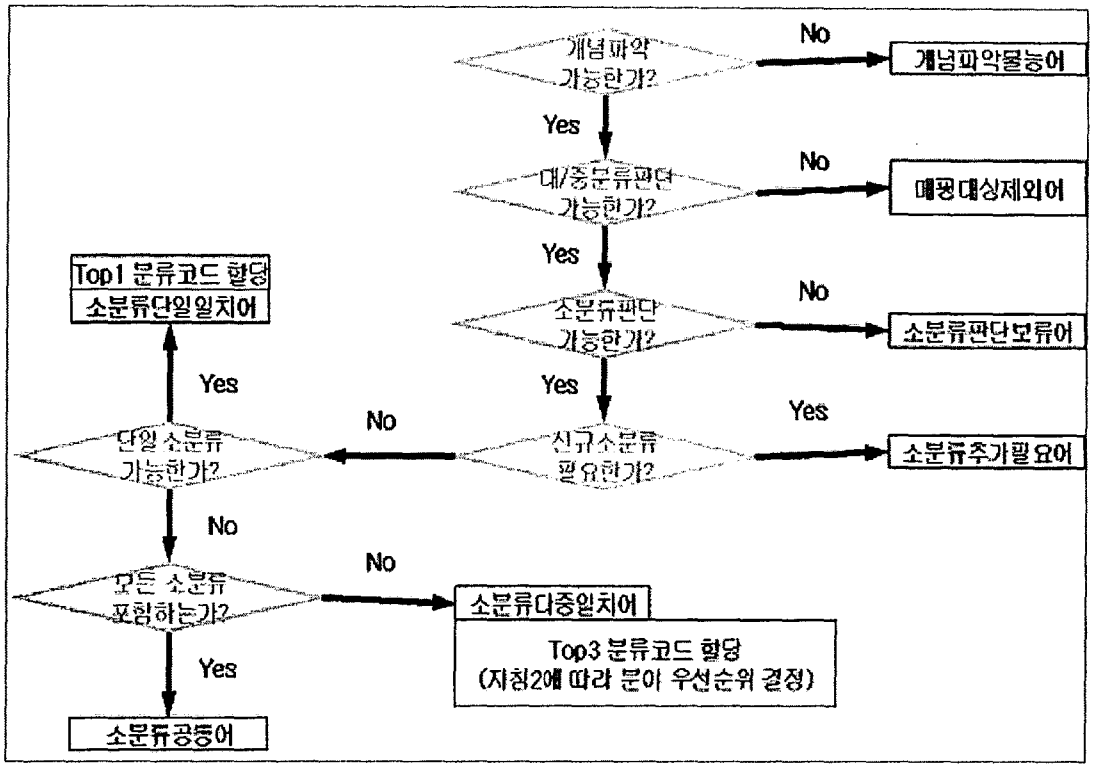

그림 2. 시소러스-분야분류체계 간 매핑 Workflow

표 3은 전체 15,000여 개의 개념어 중 우선적으로 선정한 4,579개 개념어에 대한 매핑 결과를 보여준다. '소분류 단일 일치어', '소분류 다중 일치어', '소분류 공통어', '소분류 판단 보류어', '소분류 추가 필요어'는 개념어에 분야분류명을 하나 또는 복수 개 할당할 수 있는 경우이며, 나머지 두 유형은 분야분류명 할당이 불가능한 경우이다. 즉, 76.57%에 해당하는 개념어들에는 분야분류명을 할당하는 것이 가능하며, 이는 시소러스 개념어를 매개체로 하여 문헌에 분야를 할당하는 것이 현실적으로 가능함을 의미한다.

표 3. 4,579개 개념어에 대한 분야분류명 매핑 결과

매핑 유형	개수	비율
소분류 단일 일치어	2,105	45.79%
소분류 다중 일치어	563	12.25%
소분류 공통어	332	7.22%
소분류 판단 보류어	521	11.33%
소분류 추가 필요어	7	0.15%

개념파악 불능어	30	0.65%
매핑대상 제외어	1,039	22.60%
합계	4,597%	100%

4.4 주제 및 분야할당 시스템

그림 3은 정보검색 시스템에서의 색인기와 시소러스, 분야분류체계를 결합하여 주제 및 분야를 할당하는 시스템의 구성도를 보여준다. 본 시스템은 문헌으로부터 색인어를 추출하는 과정, 색인어와 시소러스 개념어를 매칭하는 과정, 매칭에 성공한 개념어들과 이들에 부착된 분야분류명을 이용하여 주제 및 분야를 결정하는 과정, 해당 문헌에 주제 및 분야를 할당하는 과정으로 나누어볼 수 있다.

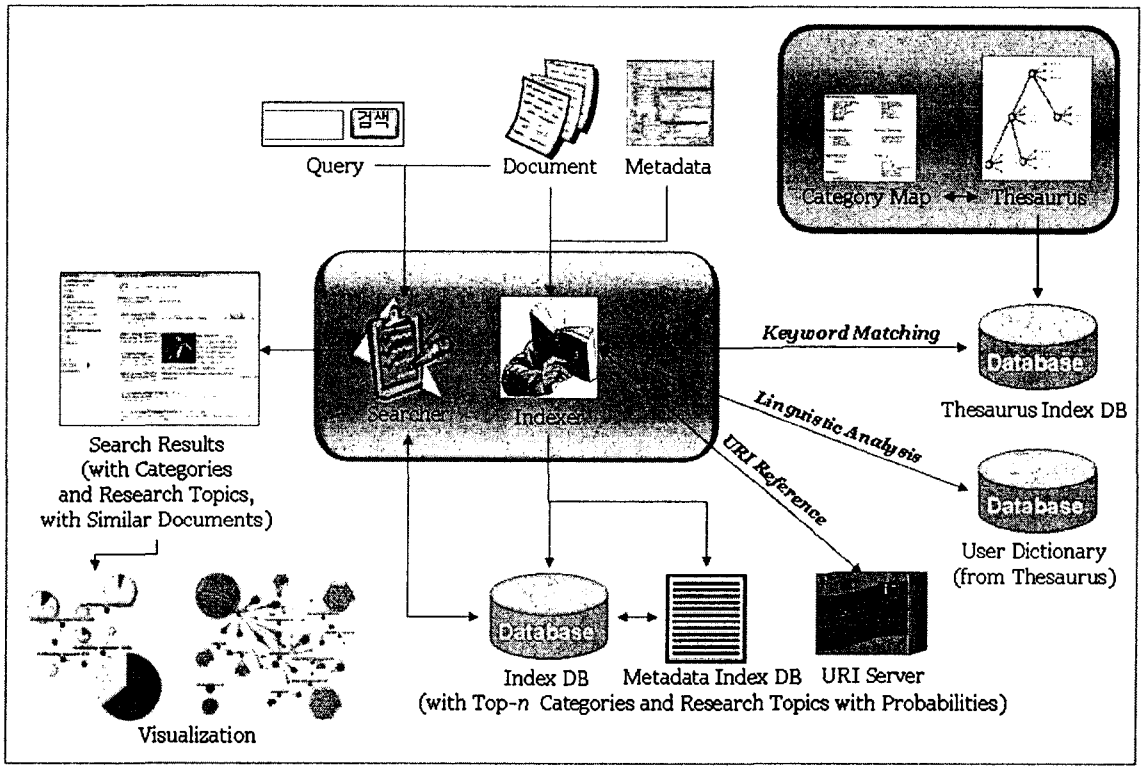

그림 3. 정보검색 시스템을 이용한 주제 및 분야할당 시스템 구성도

(1) 색인어 추출: 일반적인 정보검색 시스템의 색인기를 이용하여 문헌 내 텍스트를 언어분석하고 명사 위주의 색인어를 추출하는 과정이다. 이때, 유효한 색인어가 추출될 수 있도록 시소러스 개념명들을 사용자 사전에 반영하는 작업이 필요하다.

(2) 색인어-개념어 매칭: 추출된 색인어들은 문헌을 대표하는 키워드들이라고 할 수 있다. 이들과 시소러스 개념어들 간의 매칭을 통해 통제된 키워드들을 선별한다. 본 시소러스는 범용 과학기술 시소러스로서 과학기술 분야를 대표할 수 있는 개념어들로 구성된다. 키워드 통제를 엄격하게 하기 위해 불용어 사전을 별도로 두어 주제로서 부적절한 개념어들을 미리 배제한다. 불용어 사전을

구축하기 위해 먼저 15,000여 시소러스 개념어들 중 7,175건의 과학기술문헌[5]에 출현한 4,472개의 개념어들을 추출하였다 (정한민 외 2006) (황순희, 정한민, 성원경 2005) (Jung, Sung, and Park 2006). 이들을 DF (Document Frequency)를 기준으로 정렬하고 DF값이 큰 개념어들을 우선적으로 검토하여 불용어 사전에 추가하였다. DF값이 크다는 것은 많은 문헌에서 골고루 출현한다는 의미로 문헌 간 변별력을 떨어뜨리므로 정보량이 작다고 볼 수 있다. DF값의 분포를 살펴보면 그림 4와 같다. 대부분의 개념어들은 낮은 DF값을 가지며, 불용어 관점에서 본다면 높은 DF값을 가진 개념어들 (예. "정보", "방법", "시스템", "데이터" 등)과 아주 낮은 DF 값을 가진 개념어들 (예. "동맥경화증", "손가락뼈", "도시가스", "가스보일러" 등)이 불용어 처리되었다. 아주 낮은 DF값을 가진 개념어들은 주제로서 부적절한 수준의 사용이 제한되거나 범용적이기 때문에 불용어 처리된 경우가 많다. 본 과정을 통해 최종적으로 1,946개 개념어를 불용어 처리하였다.

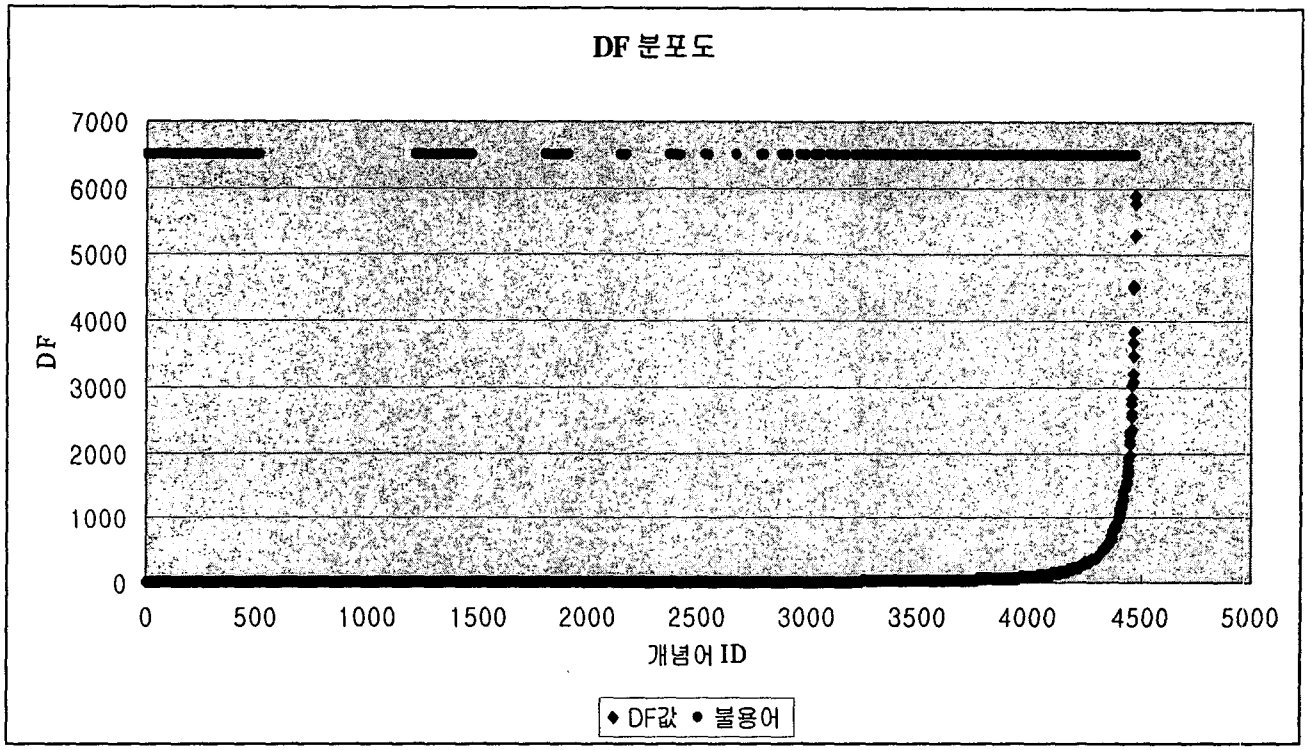

그림 4. DF 관점에서의 시소러스 개념어와 불용어 간의 상관관계

(3) 주제 및 분야 결정: 상기 과정들을 통해 얻은 개념어와 매칭된 색인어들을 이용하여 해당 문헌의 주제 및 분야를 결정한다. 이 과정에서는 색인어 목록, 개념어 목록, 색인어와 매칭된 개념어 목록, 색인어 별 TF (Term Frequency), 색인어와 매칭된 개념어 별 TF, 색인어와 매칭된 개념어 별 시소러스에서의 깊이, 색인어와 매칭된 개념어 별 개념패싯 (Conceptual Facet), 색인어와 매칭된 각 개념어에 부착된 분야분류명 목록 등의 정보를 사용한다. 자세한 주제 및 분야할당 알고리즘은 4.5절에서 기술한다.

(4) 주제 및 분야할당: (3)을 통해 획득한 주제 및 분야 별 상위 n개의 목록과 각각의 확률값들은 해당

[5] 2002년부터 2006년까지의 한국정보과학회, 대한전자공학회, HCI학회, 한국정보처리학회 학술대회논문집이 대상

문헌에 동적 할당된다. 추후 시소러스와 분야분류체계가 변경되는 경우 전체색인을 통해 (1) ~ (4) 과정을 반복하면 자동적으로 문헌들에 대한 주제 및 분야가 갱신된다. 이 자동화 과정은 콘텐츠와 정보가 강하게 결합하여 발생하는 유지보수 및 확장 문제를 해결함으로써 지속적인 정보확장을 가능하게 한다는 장점을 가진다.

4.5 주제 및 분야할당 알고리즘

본 알고리즘은 문헌에 해당 문헌을 대표하는 주제 및 분야를 할당하기 위한 것으로, 확률값을 가지고 중요도 순으로 주제 및 분야를 순위화한다. 이를 위해 주제로 사용될 시소러스 개념어들의 전체 색인어 집합에 대한 포함 범위가 일정 수준 이상일 것이라는 가정 하에 개념어와 매칭된 색인어들을 대상으로 통제된 주제와 분야를 제시한다. 4.4절 (3)에서 언급한 바와 같이 본 알고리즘을 위해서는 다양한 정보들을 사용한다. 이들을 정규적 형태로 기술하면 다음과 같다.

(1) 색인어 목록: 전체 문헌 집합 중 k번째 문헌 $D_k = \{t_{k1},...,t_{km}\}$ 는 m개의 색인어를 가진다. t_{ki} 는 D_k 에 나타난 i번째 색인어를 의미한다.

(2) 개념어 목록: 시소러스 개념어 집합 $S = \{s_1,...,s_p\}$ 는 전체 p개의 개념어를 가진다.

(3) 색인어 별 TF: $tf_{D_k}(t)$ 는 문헌 D_k 내에서 용어 t가 출현한 빈도수 (TF)이다.

색인어와 매칭된 개념어 별 TF: $tf_{D_k^S}(s)$ 는 색인어와 매칭된 개념어 빈도수이며, 이것은 한 문서 내에서 해당 개념어 s로 대응되는 색인어들의 빈도수의 합이다. 즉, $tf_{D_k^S}(s) = \sum_{\substack{s=f_{\text{동의어그룹대표어}}(t),\\ t\in D_k \,\mathrm{I}\, S}} tf_{D_k}(t)$ 이다.

여기에서 주목할 내용은 시소러스 상에서의 동등관계 ('USE/UF 관계'와 'RT-동등관계')에 해당하는 개념어들 (동등개념어 집합)은 대표개념어로 정규화된다는 것이다. 대표개념어는 일반적으로 USE를 사용하며, 일단 정해지면 동등개념어 그룹 내에서 USE가 변경이 되더라도 전체색인을 수행하기 전까지는 대표개념어를 유지해야 한다. 그렇지 않으면, 동등개념어 집합 내에서의 서로 다른 개념어들이 주제로서 문서에 할당되는 혼란이 일어날 수 있다. $f_{\text{동의어그룹대표어}}(s)$ 는 개념어 s가 속한 동등개념어 집합 내의 대표개념어를 반환하는 함수이다. D_k^S 는 D_k 의 색인어들 중 시소러스 개념어에 매칭되는 색인어들을 그 색인어와 매칭된 동등개념어 집합의 대표개념어로 정규화한 색인어 집합으로 $D_k^S = \{f_{\text{동의어그룹대표어}}(t) \mid t \in D_k \,\mathrm{I}\, S\}$ 에 해당한다.

(4) 색인어와 매칭된 개념어 별 TF: 각 개념어 s에 대해 문헌 D_k 내에서의 개념어 빈도수는 $tf_{D_k^S}(s)$ 와 같이 정의한다.

(5) 색인어와 매칭된 개념어 별 시소러스에서의 깊이

(6) 색인어와 매칭된 개념어 별 개념패싯 (Conceptual Facet)

(7) 색인어와 매칭된 각 개념어에 부착된 분야분류명 목록

상기 정보를 이용하여 주제 및 분야를 결정하고 문헌에 할당하는 예를 다음 단락을 예로 들어 설명하면 다음과 같다.

> ...
> 휴대폰 운영체제 시장을 노리고 있는 MS의 경우 이미 지난해 10월 개발자 컨퍼런스에서 윈도우 모바일 개발 플랫폼을 확장하여 컴퓨터 내에서 정사각형 스크린, portrait, VGA 등을 포함한 스크린 화상과 음성인식 기능을 지원할 것임을 선언했다. 이에 발맞춰 대안 세력으로 떠오르고 있는 임베디드 리눅스와 심비안 계열의 기기들 또한 현재의 VGA, QVGA를 넘어서 소형 노트북 수준의 해상도를 지원할 수 있는 단말기를 내놓을 계획을 발표했다. CDMA의 경우 이미 퀄컴 MSM6100 이상의 칩셋에서는 브루(BREW) 기반의 3D 게임 환경이 자리잡아가는 상황인 만큼 10년 뒤라면 모바일폰을 비롯한 휴대폰의 그래픽은 3D 게임 환경을 넘어서 홀로그램 기반의 3차원 게임이 등장할 가능성도 전혀 먼 세계의 소설 줄거리는 아닐 것이다.
> ...

(1) 색인어 목록: D_k = {"컴퓨터", "휴대폰", "단말기", "모바일폰", "음성인식"} (더 많은 색인어들이 실제 존재하나 본 예제에서는 생략한다.)

(2) 개념어 목록: S = {"단말기", "휴대폰", "핸드폰", "휴대전화", "모바일폰", ...}. {"휴대폰", "핸드폰", "휴대전화", "모바일폰"}은 동등개념어 집합이며 ("휴대전화"와 "휴대폰", "핸드폰"은 'USE/UF 관계'이며, "휴대전화"와 "모바일폰"은 'RT-동등관계'이다.), "휴대전화"를 대표어로 정의했다고 가정한다.

(3) 색인어 별 TF: D_k^S ={"단말기", "휴대전화"} ($f_{\text{동의어그룹대표어}}$ ("단말기") = "단말기", $f_{\text{동의어그룹대표어}}$("휴대폰") = "휴대전화", $f_{\text{동의어그룹대표어}}$("모바일폰") = "휴대전화"에 의해 색인어와 매칭된 동등개념어 집합의 대표개념어로 정규화한 색인어 집합이 결정된 것이다.). tf_{D_k}("컴퓨터") = 1, tf_{D_k}("휴대폰") = 2, tf_{D_k}("단말기") = 1, tf_{D_k}("모바일폰") = 1, tf_{D_k}("음성인식") = 1

(4) 색인어와 매칭된 개념어 별 TF: $D_k \cap S$ ={"컴퓨터", "휴대폰", "핸드폰"}, $tf_{D_k^S}$ ("단말기") = 1, $tf_{D_k^S}$("휴대전화") = tf_{D_k}("휴대폰") + tf_{D_k}("모바일폰") = 3

(5) 색인어와 매칭된 개념어 별 시소러스에서의 깊이: Depth("단말기") = 1, Depth("휴대전화") = 3

(6) 색인어와 매칭된 개념어 별 개념패싯 (Conceptual Facet): CF("단말기") = '기기·장치·부속', CF("휴대전화") = '기기·장치·부속'

(7) 색인어와 매칭된 각 개념어에 부착된 분야분류명 목록: Theme("단말기") = {'컴퓨터공학', '전기 및 전자공학'}, Theme("휴대전화")={'전기 및 전자공학'}

주제: 휴대전화 (75%), 단말기 (25%)
분야: 전기 및 전자공학 (66%), 컴퓨터공학 (33%)

* 본 예제에서는 상위 3개까지의 주제 및 분야를 결정하고 이들의 전체 TF 합 (주제의 경우 1 + 3 = 4, 분야의 경우 2 + 1 = 3)으로 각 TF를 나누는 방식으로 확률값을 결정한 결과를 보여준다.

5. 실험

현재 본 연구에서 제시한 주제 및 분야할당은 정보유통 서비스에서 문헌을 등록할 때 원본으로부터 자동으로 문서 필터링을 하여 텍스트를 추출한 후 4장에서 제시한 방법을 이용하여 이루어진다. 그림 5는 정보유통 서비스에서 논문에 대한 메타데이터와 원문을 등록하고 등록결과를 보여주는 인터페이스 예이다. 등록 과정에서 주제 및 분야는 자동으로 추출되어 해당 문헌에 동적으로 할당된다.

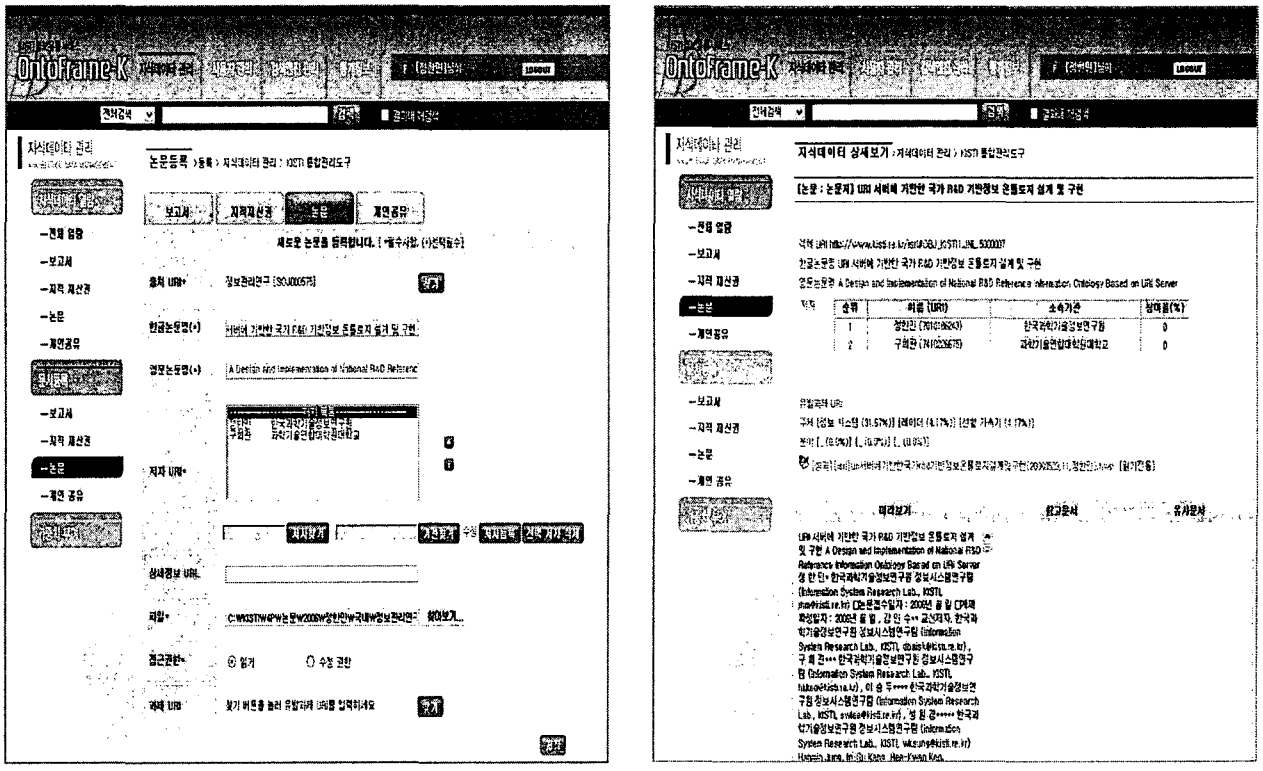

그림 5. 정보유통 서비스에서의 문헌 등록화면 및 등록결과 예

그림 6은 주제 및 분야를 한정하고 특정 성과정보를 제시하거나 연구자의 연구영역을 파악할 수 있는 등의 추론결과를 활용할 수 있는 추론 서비스 화면 예를 보여준다. 그림 7은 추론규칙을 이용하여 문헌에 주제 및 분야를 온톨로지 기반으로 할당하는 예를 보여준다. 그림 5에서 동적으로 할당된 주제나 분야는 그림 7에서의 예와 같은 추론규칙을 이용하여 추론엔진이 직접적으로 이용할 수 있는 형태로 구체화된다.

국가 과학기술 R&D 기반정보 온톨로지에서는 학술논문의 경우 최대 3개까지의 주제를 가질 수 있도록 정의하고 있으며, 각 주제 별로 주제가중치를 부여할 수 있다. 주제가중치는 5장의 주제 및 분야할당

알고리즘의 예에서 살펴 본 확률값에 대응하는 것으로 0에서 1사이로 정규화한 값이다. 각 주제영역은 주제키워드를 가지는 데 이것이 시소러스 개념어에 해당한다. 본 온톨로지는 URI 서버와 연계된 형태로 구현된 최초의 온톨로지로서 각 개념어는 주제 URI로 표현된다 (강인수 외 2006) (정한민 외 2006). 학술논문이 해당 주제들로 분류된다는 정보는 온톨로지에 표현되어 있지 않다. 그렇지만, 추론규칙을 이용하여 ' isClassifiedBy' 관계 (Property)를 정의하고 학술논문 객체와 주제키워드를 이 관계로 직접 연결하면 다음과 같은 RDQL (RDF Data Query Language)을 이용하여 쉽게 특정 주제의 문헌들을 저술한 연구자들을 검색할 수 있다.

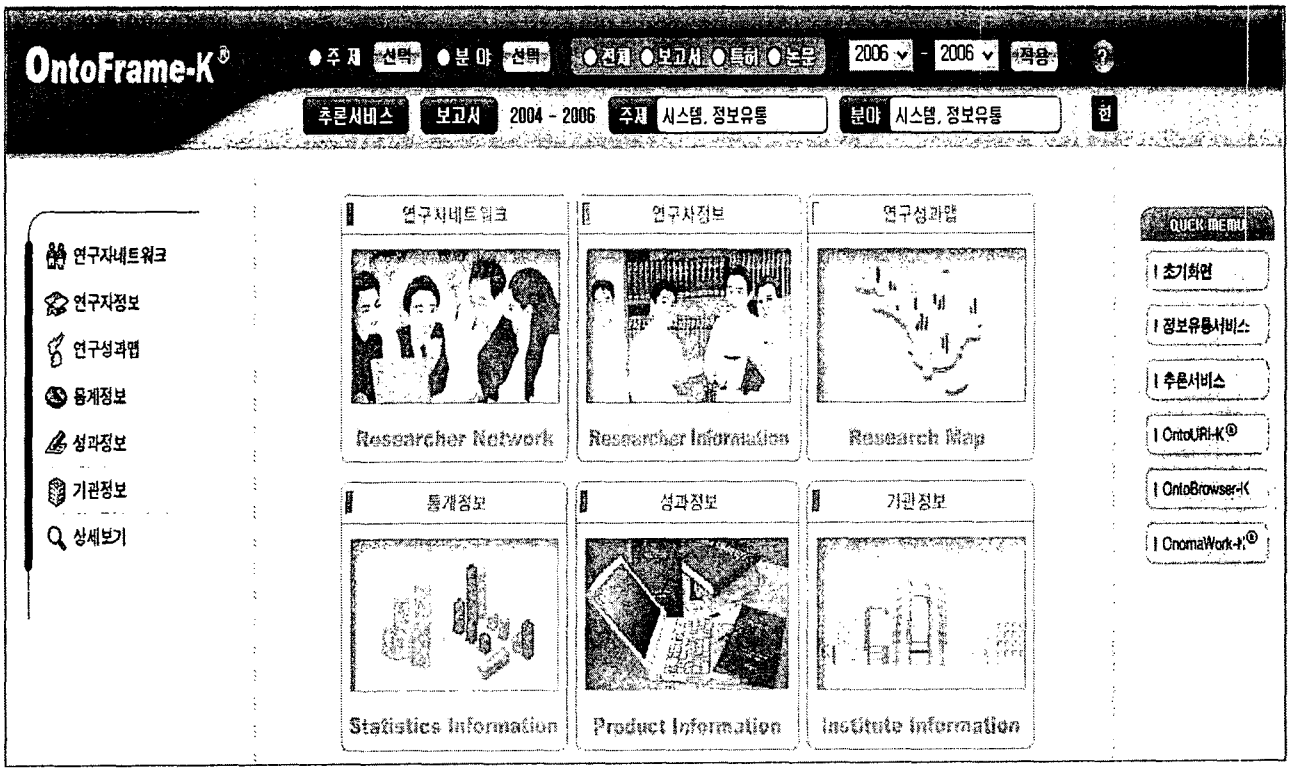

그림 6. 추론 서비스 화면 예

```
SELECT ?y
WHERE (?x wasCreatedBy ?y) (?x isClassifiedBy <http://www.kisti.re.kr/isrl#TOP_040305>)
```

* ?x와 ?y는 변수이다. 본 RDQL 예에서 ?x는 성과정보를, ?y는 연구자를 의미한다. 이 질의는 주제 URI로서 <http://www.kisti.re.kr/isrl#TOP_040305>을 가지는 성과정보 ?x를 먼저 찾고, ?x를 저술한 모든 연구자 ?y를 검색하고자 하는 것으로, 추론엔진은 일반적으로 Backward Chaining 방식으로 추론을 수행하여 원하는 결과를 제시한다.

6. 결론

본 연구는 시소러스, 분야분류체계를 이용하여 과학기술문헌에 동적으로 확률값과 함께 주제 및 분야 정보를

할당하는 기법을 설명하였다. 본 연구가 기존 많은 연구들이 콘텐츠와 관련정보를 수작업을 통해 강하게 결합시키거나 통제되지 못한 주제어를 사용하여 일관성 있는 정보확장이나 서비스 품질 향상에 제대로 대처할 수 없었던 문제들을 해결함으로써 정보의 지속적인 확장 및 유지보수를 용이하게 할 수 있다는 장점을 제공한다. 또한, 정보유통 서비스에서 과학기술문헌들을 주제 별, 분야 별 등 다양한 관점으로 분류함으로써 특정 문헌과 관련있는 문헌들을 쉽게 검색할 수 있도록 도와준다. 본 연구를 통해 구현된 주제 및 분야할당 시스템은 중복문헌 검사, 디렉터리 서비스, 유사연구 검색 등 다양한 응용에서 활용될 수 있을 것으로 기대한다.

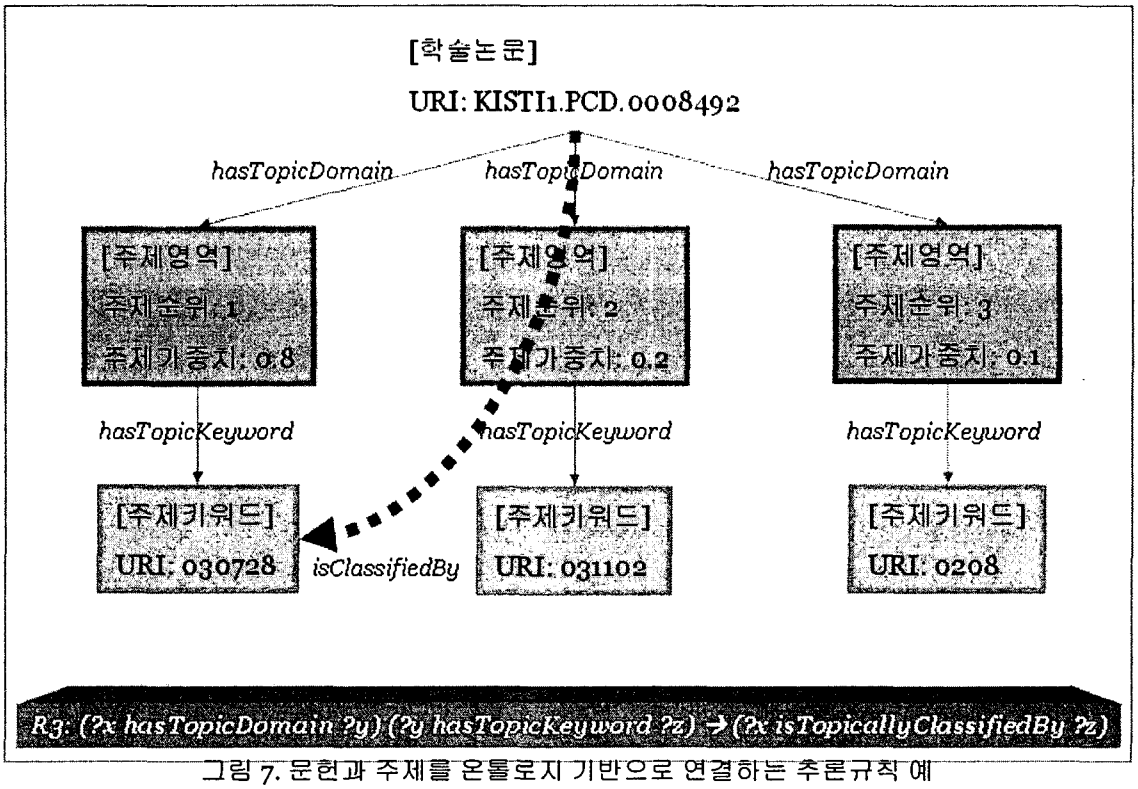

그림 7. 문헌과 주제를 온톨로지 기반으로 연결하는 추론규칙 예

참고문헌

강인수, 정한민, 이승우, 김평, 성원경, 2006. 국가 과학기술 R&D 기반정보 온톨로지. 한국콘텐츠학회 제4회 춘계종합학술대회논문집.

방선이, 양재동, 양형정, 2004. k-NN 분류 알고리즘과 객체 기반 시소러스를 이용한 자동 문서 분류. 정보과학회논문지: 소프트웨어 및 응용 31 (9).

안찬민, 박선, 박상호, 최범기, 이주홍, 2004. 분류 주제 자동 생성 및 동적분류체계 방법을 이용한 이메일 분류. 한국정보과학회 춘계학술대회논문집.

이용배, 맹성현, 2003. 장르와 주제 범주간 용어 편차정보를 이용한 디지털 문서의 장르기반 분류. 정보과학회논문지: 소프트웨어 및 응용 30 (1).

이창범, 박혁로, 2001. 시소러스를 이용한 문서 자동 요약. 한국정보과학회 춘계학술대회논문집.

정한민, 이승우, 강인수, 성원경, 2006. 온톨로지 구축 지원을 위한 과학기술 문헌으로부터의 인력정보 구축.

한국콘텐츠학회 제4회 춘계종합학술대회논문집.

정호석, 임종태, 나혜숙, 민철호, 2000. 자동 문서 분류를 위한 분류 주제어의 자동 증식 방법. 한국정보과학회 추계학술대회논문집.

황순희, 정한민, 성원경, 2005. 어휘의미망의 이론과 실제 - 과학기술 분야 전문용어 구축과 활용을 중심으로 -. 한국어의미학회 제17회 학술발표회논문집.

Jung, Hanmin, Sung, Won-Kyung, and Park, Dong-In, 2006. Project Report on a Korean Science & Technology Thesaurus with Conceptual/Relational Facets. Proceedings of the 3rd Global WordNet Conference.