특허와 학술문헌 강결합 연계를 위한 프레임웍 개발

Development Framework for Tightly Coupled Linking of Patent and Scientific Paper

노경란, 김완종, 권오진, 서진이
한국과학기술정보연구원 정보융합개발팀

Noh Kyung-Ran, Kim Wan-Jong, Kwon Oh-Jin, Seo Jinny
KISTI, Information Fusion Development Team

요약

정보의 폭발적인 증가로 인해 연구 개발을 위한 전 과정 중 연구동향 분석에 많은 시간이 소모되고 있다. 최근 특정 분야의 지식이 연구개발이나 제품개발로 이루어지던 시대에서 융합지식을 통한 연구개발이나 제품생산으로 빠르게 진화하고 있다. 이러한 패러다임을 수용하기 위해 기존의 독립적이고 단편적인 정보로부터 융합정보를 제공할 수 있는 체계로의 전환이 필요하게 되었다. 또한 과학 기술 정책 및 산업 정책을 수립하기 위해 최근 과학, 기술, 산업의 지식 흐름에 대한 연구가 활발히 진행되고 있으나 정량적인 분석을 활용하기란 매우 어려운 문제이다. 왜냐하면 과학-기술간 지식흐름을 분석할 수 있는 정보자원이 존재하지 않기 때문이다. 이 연구는 연구개발이나 과학기술정책 및 산업정책에 활용할 수 있는 특허정보와 학술 문헌간 강 결합 연계 체제를 갖는 프레임웍을 개발하고자 한다.

Abstract

Because of explosive increase of information, it spends a lot of time to trace and analysis research trends during total R&D process. It has rapidly evolved from R&D or process development within a specific domain of knowledge to R&D or process development through knowlede convergency. To accept such a paradigm, it is necessary to convert dissemination system from a separate, standalone, and fragmentary information to highly coulpled fusion information. Although there are several studies on knowledge flows between science and technology or technology and industry, it is difficult to analyse and utilize quantitatively to establish policy of Science, Technology, and Industry. The reason is the lack of information resource to analyse knowledge flow from science to industry. This paper intends to develop framework of highly coupled linking system between patent and scientific paper to utilize R&D, S&T policy, and industry policy.

1. 서 론

과학기술간 관계가 국가경제발전에 미치는 영향이 실증적으로 입증됨에 따라 과학과 기술간 연계를 찾고자 하는 연구들이 경제학적 측면에서 진행되었다.

디지털 시대의 도래와 함께 데이터 전산화가 훨씬 쉽고 신속해짐에 따라 1990년대 후반부터 과학과 기술간 관계를 특허에 인용된 비특허 문헌을 통해 측정하고자 한 연구들이 있었다.

2000년대 들어 디지털 환경이 확산되면서 콘텐트를 둘러싼 경쟁은 날로 격화되고 있다(전자신문 2006.9.25). 특히 디지털 컨버전스의 진행과 함께 새로운 콘텐트 서비스가 등장하면서 융합이 최대 이슈로 떠오르고 있다. 그리고 서로 다른 유형의 콘텐트를 결합시키기 위한 노력이 대규모 콘텐트 제공업체를 통해 이루어지고 있다. 디지털 컨버전스 시대에 맞춤형, 쌍방향으로 변화하는 미디어 조류에 적극적으로 대처하기 위해 원소스 멀티유스(One Source Multi Use)의 극대화 전략이 예외없이 중시되고 있다. 다단계로 가치를 부가하는 원소스 멀티유스 전략은 디지털 콘텐트의 중요성과 비중이 높아지고 있는 요즘 더욱 명확해지고 있다. 하나의 성공적인 콘텐트를 다양한 활용을 통해 원소스 멀티유스를 잘 실행해내고 부가가치를 부여하는 것이 필요하다.

이 논문은 지금까지 서로 분리되어 독자적으로 다루어졌던 특허와 과학논문이라는 콘텐트를 하나의 융합콘텐트로 결합시키기 위해 특허에 인용된 비특허문헌을 이용하여 융합 프레임웍을 제시하고자 한다. 기존의 특허와 과학논문간 연계모형이 지니고 있는 여러 제한점들로부터 진보된 모형을 제시하고자 한다.

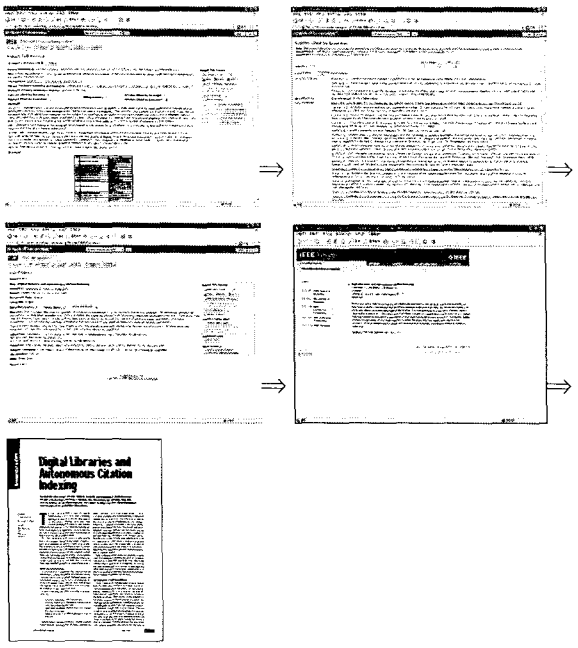
2. 기존 사례

2.1 Derwent Innovation Index

특허정보는 경쟁사의 활동을 모니터링하고 최신기술을 파

악하며, 연구개발자원을 절감시키며, 새로운 아이디어의 영감을 얻게 한다. 특허에 관심을 두는 것은 기술콘텐트의 내용 파악이나 사업 기획 및 개발을 위해 또는 기관내 발명을 보호하기위한 목적이 있다.[1].

Thomson Scientific 은 특허와 인용정보를 검색할 수 있도록 Derwent Innovation Index[SM]를 서비스하고 있다. DII는 특허 심사관과 발명자가 인용한 특허 및 문헌에 대한 정보를 제공한다.



▶▶ 그림 1. DII의 특허-문헌 연계

DII는 특허에 인용된 과학논문에 대한 서지정보뿐만 아니라 WOS 로의 링크를 통해 원문을 제공한다. DII는 피인용문헌(cited references)에 대한 정보를 제공함으로써 현재 특허의 선행기술정보 혹은 특허의 배경이나 전개 및 중요성 등을 파악할 수 있도록 한다. 그러나 DII는 특허로부터 과학논문으로 링크되는 단반향만 체제를 가지고 있으며 과학논문에서 특허로의 링크는 이루어지지 않는 한계점을 가지고 있다.

2.2 과학논문정보 추출

Verbeek 등(2002)은 과학과 기술간 연계구조를 분석하기 위해 특허에 인용된 과학논문을 이용하였다. 과학논문을 식별하기 위해 인용정보의 텍스트구조에 중점을 둔 키워드-결합 탐색알고리즘(keyword- assembled search algorithms)에 기반한 프로그램을 개발하였다. 예를들면 키워드'journal'과 이의 변형인 J., Jr., Journal 등과 같은 단어가 포함된 인용정보를 추출한 후, 키워드 volume이 포함된 인용정보를 추출하였으며, encyclopaedia, dictionary, manual 등과 같은 키워드가 포함된 인용정보를 삭제하였다. 세단계의 추출과정을 거쳐 얻어진 과학논문 인용정보에 파싱알고리즘을 적용한 후 SCI 과학논문과 매칭하였다.

Tamada 등(2004)은 과학이 특허기술변화에 끼친 영향을 발견하고자 일본특허에 인용된 과학논문을 추출하였다. 비특허문헌을 추출하기 위해 6가지 유형의 규칙을 발견하였다. 즉 "Japanese index", 2가지 유형의 "English index", volume과 page처럼 숫자와 기호가 결합된 2가지 유형, 그리고 서기연도를 적용하여 비특허문헌을 추출하였다.

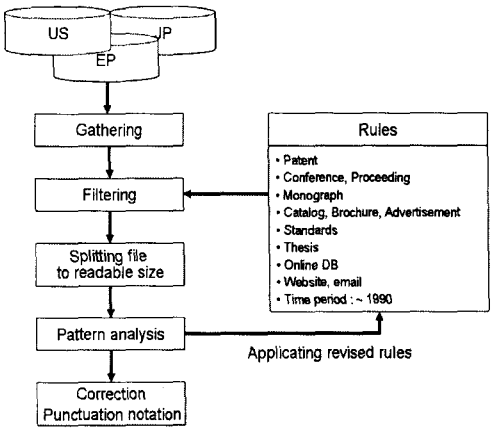
3. 학술문헌과 특허 연계를 위한 작업

본 연구는 과학논문과 특허간 연계를 위해 특허에 인용된 과학논문을 이용한다. 특허에 인용된 과학논문은 미국특허의 경우 프론트페이지에 "other reference"라는 항목아래 비특허문헌으로 기재된다. 비특허문헌에는 학술커뮤니티에서 가장 기본적인 정보유통수단인 학술지 또는 학술회의자료에 수록된 과학논문뿐만 아니라 단행본, 기술보고서, 산업표준 및 규격, 매뉴얼, 출원특허 등으로 구성된다.

학술문헌과 특허를 연계하기 위해 3단계의 모듈을 개발하였다. 먼저 과학 논문 추출 단계, 두 번째, 과학논문 정형화 단계, 세 번째 과학논문 매칭을 통한 서지정보 표준화단계로 구성된다.

3.1 제1단계 과학논문 추출

비특허 문헌중에서 과학논문을 추출하는 작업은 3단계의 작업은 가장 많은 시간이 소요되는 작업이다.



▶▶ 그림 2. 특허에서 과학논문 추출모듈

인용논문을 추출하는 것이 어려운 이유는 심사관이나 발명자의 서지기술방식이 일관성이 결여되어 있기 때문이다. 따라

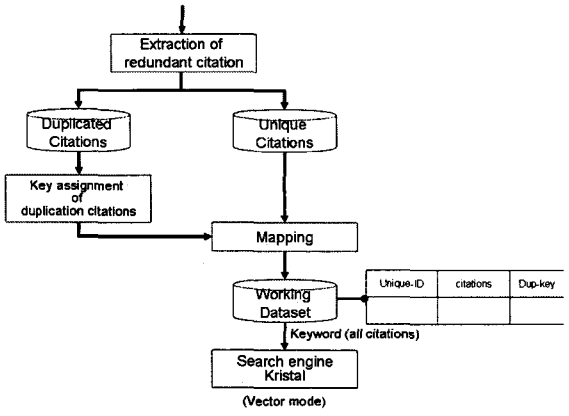
1) 톰슨웹사이트

서 동일한 유형의 자료임에도 불구하고 여러 가지 패턴으로 기입되는 규칙을 찾아내는 것은 시간소모적인 작업이다.

비특허문헌중 과학논문을 추출하기 위해 과학논문 이외의 유형을 제거하는 방법을 사용하였다. 예를 들면 다수를 차지하고 있는 출원특허와 관련된 키워드(patent, inventor, application, JPO 등), 학술회의자료와 관련된 키워드(conference, workshop, symposium, seminar, congress, meeting 등), 기업의 제품홍보자료와 관련된 키워드(catalog, brochure, advertisement, guidebook, databook 등), 규격과 관련된 키워드(ASTM, DIN, ETSI 등), 웹사이트와 관련된 키워드(http://)를 키워드-결합 탐색알고리즘을 사용하여 작업 대상 집합의 크기를 줄여나가는 방법을 사용하였다. 작업 대상 집합을 줄인 후 예외 사항에 대한 필터링 규칙을 보완하여 이를 반복 수행하였다. 다음 단계로 중복 레코드에 대한 불필요한 연계작업을 줄이기 위해 구두점 보정작업을 수행하였다. 단계1을 통해 최소의 노이즈를 갖는 작업 대상 집합을 구하였다.

3.2 제2단계 과학논문 정형화

파싱알고리즘의 성공은 과학논문을 식별하는 과정에서 가장 핵심적인 단계이다. 특허에서 비특허문헌의 인용여부를 결정하는 특허심사관은 기존에 인용된 문헌의 서지정보를 그대로 인용하는 성향을 가지고 있다. 이에 따라 동일형태로 중복 기술된 서지정보에 대해 불필요한 연계작업을 줄이기 위해 동일한 레코드에 대한 중복키를 부여하였다. 중복된 키를 가지고 있는 테이블과 매핑 작업을 통해 중복키를 포함하는 전체 작업 대상 테이블을 구성한다.

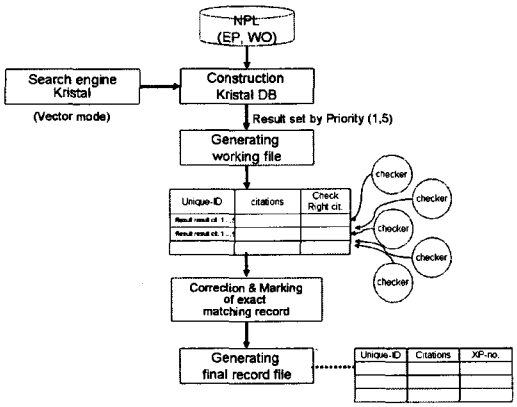


▶▶ 그림 3. 과학논문 파싱 - 정규화 모듈

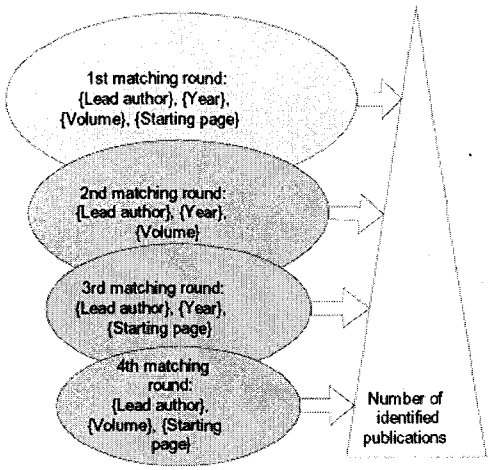
3.3 제3단계 과학논문 매칭 및 서지정보 표준화

비특허문헌에 인용된 과학논문에 포함된 키워드들을 검색어로 사용하여 학술정보 데이터베이스와 매칭작업을 수행한다. 비특허문헌에서 추출한 과학논문에 대한 서지정보는 학술지명, 기사명, 저자명 등의 정보중 하나가 누락된 경우가 많다. 먼저 제1저자명, 발행년, 권정보, 시작페이지를 모두 사용하고, 그다음 제1저자명, 발행년, 권정보만 이용하는 것과 같이 단계별로 매칭요소를 변경하여 매칭작업을 수행한다.

서지데이터베이스와 매핑결과 복수개의 서지정보가 나왔을 경우 육안으로 검증 후 최종 데이터파일을 구성한다.



▶▶ 그림 4. 과학논문 서지정보 매칭모듈



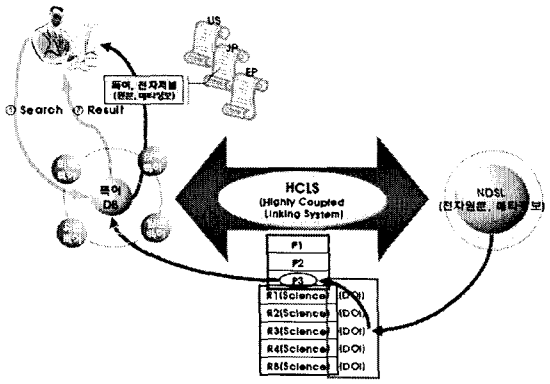
▶▶ 그림 5. 서지정보의 질의에 따른 검색결과 크기

4. 강결합 프레임 웍

본 연구에서는 3장에서 구축된 학술문헌 인용정보를 이용하여 DII의 단점인 단방향성 링크 특허→과학논문간 링크를 특허↔과학논문간 쌍방향적 링크가 가능하도록 연계모형을 개발하였다.

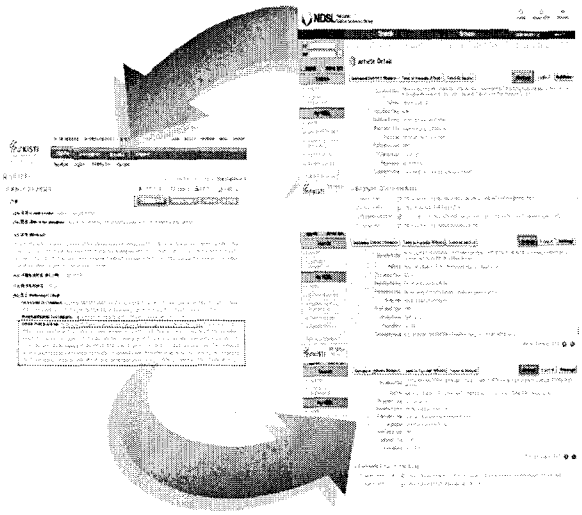
즉, 특허를 통해 표현된 기술을 개발하는데 과학적 기반을 마련한 과학논문 인용DB와 독자적으로 운영되고 있는 학술지 기사DB를 연계할 수 있는 강결합 링킹시스템 (Highly Coupled Linking System, HCLS)을 개발하였다. HCLS 는

특허와 문헌을 연계하는 솔루션으로 특허에서 문헌으로 가는 정보, 문헌에서 특허로 가는 정보를 담고 있다.



▶▶ 그림 6. 특허-문헌간 강결합 연계시스템

이용자는 특허 DB를 검색한 후 이 발명특허의 개발과 관련된 학술정보에 대한 상세한 메타정보뿐만 아니라 전자원문까지도 접근할 수 있게 된다. 또한 학술정보를 검색한 후 이 학술정보와 관련있거나 활용된 기술정보인 특허로 접근할 수 있게 된다.



▶▶ 그림 7. 특허-문헌간 강결합 연계 서비스 모형

5. 결론

특허와 학술정보간 연계프레임웍은 콘텐트융합이 표방하는 원소스 멀티유즈화라는 시너지효과를 산출한다. 디지털 컨버전스 시대에는 원소스 멀티유스를 잘 실행해내고 어떻게 그 부가가치를 좀 더 돋보이게끔 하느냐가 콘텐트의 성공요건이 될 수 밖에 없다. 이런 의미에서 특허인용정보를 이용한 특허와 학술정보간 연계라는 원소스 멀티유즈화는 진정한 의미에서 콘텐트 비즈니스의 골든룰이 되고 있다.

특허심사관이 명시한 학술문헌에 대한 인용정보를 이용하여 이용자는 자신의 검색영역을 확장시킬 수 있다. 인용된 학술정보는 관련선행기술로 여겨지며, 특허가 지니고 있는 중요도와 그 발전에 대한 중요한 배경정보를 제공한다. 또한 과학기술 정책결정자뿐만 아니라 연구자들은 특정 연구분야에서 기술발전을 연구할 수 있으며, 연구개발시 중복투자를 방지할 수 있도록 한다.

▌참 고 문 헌▐

[1] Tamada, S., Naito Y., Gemba, K., Kodama, F., Suzuki, J., Goto, A. "Science linkages in technologies patented in Japan", RIETI Discussion Paper Series 04-E-034. 2004.

[2] Verbeek, A., Andries, P., Callaert, J., Debackere, K., Luwel, M., Veugelers, R., "Linking science to technology : Bibliographic References in Patents"EUR 20492/2. 2002.

[3] 한국문화콘텐츠진흥원. "디지털 사회, 해외 글로벌 미디어 기업의 융합 트렌드 ", 2004. <http://www.kocca.or.kr/_New/data/total_read.jsp>