한국특허정보의 표기법 문제 분석

A survey on the description problems in the Korean patent information

김태중, 최호남, 이명선, 신진섭
한국과학기술정보연구원

Kim tae-jung, Choi ho-nam, Lee myung-sun, Shin jin-seop
Korea Institute of Science and Technology Information

요약

특허정보는 기술정보로서 뿐만 아니라 산업정보로서 가치도 중요하게 인정되고 있다. 그러나 한국특허정보는 표기 형식과 맞춤법이 제대로 지켜지지 않고 있어 그 효용성이 떨어지며 활용하기 위해서 매우 많은 추가적인 노력이 필요하다. 이 논문에서는 일부 국내외의 데이터베이스의 평가 및 오류에 관한 연구 현황을 소개하였으며 한국과학기술정보연구원과 한국특허정보원에서 서비스하고 있는 한국특허정보에서 2008년 1월부터 10월 사이에 출원된 데이터에서 발견된 오류와 표기의 문제를 분석하고 특허정보의 가치를 높이기 위한 개선 방안을 제안한다.

Abstract

Patent information has a great value of industrial information as well as technical information. However, Korean patent information loses the effectiveness by the lack of description rule and requires much additional efforts to use them in variety. In this paper, some examples of the studies on the problems in databases were introduced and patent application(from Jan. 2008 to Oct. 2008) data retrieved from information service system of KISTI(Korea Institute of Science and Technology Information) and KIPI(Korea Institute of Patent Information) was surveyed.

Ⅰ. 서론

특허정보는 기술정보로서의 가치와 더불어 산업 동향을 분석할 수 있는 요소를 갖고 있다. 특허정보는 새로운 기술에 대한 매우 잘 분류된 기술 문헌이며 가장 최근의 내용을 담고 있는 최대의 규모를 자랑하는 보물창고이다. 특허 정보의 약 70%이상이 다른 어디에도 발표된 적이 없으며 연구개발 성과의 지표로 활용되기도 하며 특허정보의 통계학적 처리를 통한 정량적 분석은 전략 정보로 다양한 활용가치를 가지고 있다[1]. 특히 기술 및 시장 동향, 제품 등을 파악하고 기술변화 추이를 분석하기 위한 특허 맵 분석은 부문별로 활발히 연구되고 있어 KISTI의 NDSL으로부터 특허맵에 관한 문헌을 검색한 결과 170건이 검색되었으며 특허는 831건이 검색되었다.[1)]

통계학적 분석을 위해 주로 특정분야 특허의 연도별 출원동향, 출원인별 특허 점유율, 국제특허분류(IPC)별 특허 출원 동향 등이 사용되며 다양한 형태의 특허지수가 개발되어 활용되고 있다. 일반적으로 특허지수는 혁신시스템과 경제성장을 뒷받침하는 요인을 해석하는 기초 데이터, 그리고 기술분야, 국가간, 지역간, 기업간의 지식의 확산 정도를 추적하는 데이터, 특정 기술 및 산업의 구조와 발전 정도를 측정하는 데이터 등을 해석하는데 중요한 역할을 한다[2]. 특허정보의 전자화로 이러한 분석을 위한 기초 데이터의 생성은 비교적 간단한 일이 되었으나 여기에는 데이터의 정확성이 뒷받침되어야 한다. 실제 특허정보를 분석한 논문에서 검색에 사용한 질문식을 보면 “동영상 or 멀티미디어 or 멀티메

1) 한국과학기술정보연구원(KISTI)의 NDSL(http://www.ndsl.kr)에서 2008년 12월 17일에 검색한 결과

디아 or 멀티메디어 or multimedia" 등과 같이 망라적으로 데이터를 검색하기 위해 노력한 흔적을 엿볼 수 있다. 영문의 경우에 있어서도 OCR의 에러 등에 의해 'methyl'이 'rnethyl', 'methyi', 'methy1(일)' 등으로 입력되어 있고 검색도 가능함이 보고되고 있다[3][4].

이 논문에서는 한국특허를 대상으로 어떠한 오류가 있으며 어떻게 오류를 줄일 수 있을 것인가에 대해 방법을 모색해 보고자 한다. 오류를 조사하기 위해 한국특허정보원(www.kipris,or,kr)과 한국과학기술정보원의 NDSL (www.ndsl.kr)로부터 각각 특허출원일에 대해 '20080101~20081031'로 검색한 결과 33,665건과 32,433건을 대량 내려받기로 받아 분석하였다.[2)]

Ⅱ. 본론

1. 관련연구

데이터베이스에서의 오류에 관한 연구는 국내의 경우는 데이터베이스의 품질과 목록 데이터의 오류에 관한 연구가 일부 있으며 외국의 경우는 참고문헌 정보의 오류에 관한 연구와 이를 줄이기 위한 적지 않은 연구 사례가 있다. 조순영[5]은 한국교육학술정보원의 종합목록의 학위논문의 서지데이터를 중심으로 오류의 원인과 유형을 분석하였으며 기계적으로 오류를 색출하는 방안을 제시하였다. 또한, 윤정옥[6]은 한국과학기술정보연구원의 '과학기술 연속간행물 종합목록' 데이터베이스의 레코드 품질을 평가하면서 나타난 오류의 유형을 분석하고 목록 데이터베이스에서 오류를 줄이는 최선책은 목록 규칙을 숙지하고 따르는 방안이 가장 중요하다고 지적하고 있다.

참고문헌의 오류에 관한 연구는 100년 이상의 역사를 갖고 있다[7]. 전형적인 사례로 Jaroslav Hlava가 쓴 1887년 체코의 한 논문 'O Uplavici(on dysentery)'를 독일어 초록에서 저자명이 'O. Uppavici'로 소개되고 1938년에 바로잡아질 때까지 약 50년간 그대로 잘못 인용되어 왔다. 최근 인용정도로 연구성과를 평가하는 사례가 많아지면서 참고문헌의 오류에 대한 연구 또한 다양하게 이루어지고 있다. Pandit[8]는 문헌정보학 분야의 저널 5종으로부터 131건의 논문에서 1,094건의 참고문헌을 조사하여 193건의 참고문헌에서 223건의 오류를 검출하였으며 이에 대해 오류의 유형 등을 분석하였다. Lok 등[9]은 간호학 분야의 저널에 대해 참고문헌의 오류를 심각한 오류(major error)와 사소한 오류(minor error)로 나누어 통계학적인 방법으로 위험 요소를 분석하였다.

데이터베이스의 품질에 관하여 국내외에서 많은 연구가 있었으며 품질 평가 기준과 방법에 대하여도 적지 않은 제안이 있다. 관련 연구 사례 몇 가지를 소개하여 온라인 서비스로 제공되는 데이터베이스에 관하여 어떠한 기준을 요구하는가를 검토해 보고자 한다. 이들 연구들은 대부분 도서관의 서지 목록작성에 관한 내용이거나 서지 데이터베이스에 관한 내용이다. 여러 검색도구[3)]를 사용하여 조사한 결과 OCR에 의한 입력 오류이외에는 특허정보의 오류에 관한 연구를 다룬 논문을 찾을 수가 없었다. 특허정보는 서지 데이터베이스와 유사한 성격을 가지고 있다.

한국데이터베이스진흥센터의 데이터베이스 품질평가 연구[10][11]는 1995년도에 데이터베이스 표준화 사업의 일환으로 시작되었으며 데이터 품질과 서비스 품질, 일반사항에 관한 품질로 나누어 특정 데이터베이스 품질을 측정하고 평가할 수 있는 방법, 측정 도구, 절차를 제시하고 있다. 데이터 품질에 관하여는 정확성, 완전성, 최신성, 일관성 등을 서비스 품질에 대하여는 사용용이성, 사용자 지원성, 검색성, 비용, 네트워크 및 하드웨어를 그리고 일반요구사항에 대한 품질로는 수록범위, 전문성, 기타를 평가 요소로 제시하였다.

이응봉 등[12]은 선행연구 결과의 분석을 통해 데이터베이스의 품질은 데이터베이스의 바람직한 정도 또는 우수성이라고 정의하고 데이터 품질과 서비스 품질 양자를 동시에 고려하여 데이터베이스 품질 평가 기준을 제시하고 과학기술분야의 데이터베이스에 대해 품질평가를 실시하였다. 데이터 품질에 관하여는 정확성, 완전성, 최신성, 수록범위, 전문성을 그리고 서비스 품질에 대하여는 검색성/접근성, 사용 용이성, 사용자 지원성, 비용, 네트워크 및 하드웨어를 품질 기준으로 삼았다.

이제환[13]은 종합목록 DB의 품질 평가 기준을 포괄성, 배타성, 최신성, 중복성, 일관성, 완전성으로 정하

2) 2008년 12월 2일에 실용신안, 디자인을 제외한 특허 출원의 경우만을 검색

3) NDSL, NAVER, Google, Webcrawler, Engineeringvillage2 등

고 이에 대해 평가관점과 평가지표 등을 제시하고 품질 측정 결과와 품질저하 원인을 분석하고 품질 개선방안을 제안하였다.

이들 관련 연구에서 살펴본바와 같이 각 데이터베이스별 성격에 적합한 평가 항목 및 방법을 제시하고 있으며 같은 용어라 할지라도 대상으로 하는 데이터베이스에 따라 다소 다른 관점과 지표를 나타내고 있다. 그러나 보편적으로 갖추어야 할 항목으로 정확성, 일관성, 완전성, 최신성을 들고 있다. 특허 정보의 경우 입력된 데이터의 정확성을 제외한 성질에 대하여는 정해진 항목들을 일정한 기준에 따라 공개되는 정보를 제공하게 되므로 별다른 평가 대상이 될 수 없다.

2. 특허정보의 구성과 주요항목 데이터 분석

각국의 특허 출원은 법률에 규정된 절차와 방법에 따라 이루어지며 특허출원서에 포함될 내용과 정보도 법이나 절차에 의해 규정되어 있다. 일반적으로 출원을 위해 제출하는 특허정보는 서지정보와 원문정보로 구분할 수 있다. 서지정보는 발명자와 출원일 등과 같은 발명외적인 정보이며 원문정보는 발명의 상세한 설명과 특허청부범위 등과 같이 발명의 실제 내용정보로서 명세서에 해당한다[2].

서지정보의 주요 항목별로 출력하여 띄어쓰기로 구분되는 단위로 추출, 정렬하고 각 단위별로 비교를 통해 차이점을 찾아보았으며 대부분의 조사는 검색 결과가 많은 한국특허정보원의 데이터를 중심으로 수행하였다.

2.1 출원인 및 발명자의 표기 문제

출원인 및 발명자의 국적, 주소 등의 정보는 여러 형태의 분석에 다양하게 사용되므로 효과적으로 활용하기 위해서는 통일된 형식이 요구된다. 그러나 주소지가 국내의 경우는 물론이며 해외의 경우 매우 다양하게 표기되고 있으며 이에 한국특허정보원의 데이터의 경우 별도의 국가명을 기입해 이용자의 편리를 도모하였으나 이의 오류가 발견되고 있다. 예를 들면 '미국'을 '캐나다'로 '중국'을 '중공'으로 표기한 경우도 있어 이들 데이터를 활용하려면 세심한 주의와 검토가 필요하다. 국내 출원자 또는 발명자의 주소를 표기함에 있어 '대전광역시'을 예로 들면, '대전', '대전시', '대전광역시' 등으로 '충청남도'는 '충남', '충청남도' 등으로 표기되고 있어 오타를 제외하더라도 시도별 특허출원 현황을 조사하기가 쉽지 않다.

해외 출원자 및 발명자의 주소는 대체로 국내의 대리인을 통해 작성되는데 대리인별 외래어 표기의 성향을 조사해 보았으나 일정한 특징이 보이지 않는 것으로 미루어 한글 맞춤법의 외래어 표기법을 지키지 않음은 물론이거니와 자체적으로 어떠한 기준도 갖추고 있지 않은 듯하다. 표 1은 다양한 외래어 표기의 사례를 일부 보여준다. 크게 2가지 유형으로 나누어 볼 수 있다. 즉, '독일'로 표기할 것인가 '독일연방공화국' 인가의 표기형식의 문제로 국가명이나 도시명을 어떻게 그리고 어디까지 표기할 것인가와 외래어 표기법의 준수 문제이다. 외래어 표기법을 준수한다면 '동경', '도쿄', '토쿄'와 같은 혼란은 없어질 것이다.

2.2 발명의 명칭

발명의 명칭 또는 제목 부분에 있어서 크게 나누어 4가지 유형의 오류를 발견할 수 있다. 가장 쉽게 발견되는 오류는 띄어쓰기 이다. 물론 현행 맞춤법에서 전문용어의 띄어쓰기에 대해서는 어느 정도 붙여서 써도 되는 것[4]으로 인정하고 있으나 특허 출원서와 명세서에 보이는 것과 같은 하나의 문장을 전혀 띄어 쓰지 않는 것은 곤란하다. 두 번째로는 역시 외래어 표기의 문제이다. 화합물이나 특수한 분야의 경우를 제외하더라도 외래어 표기의 다양성, 모호성은 지나치게 많고 크다. 세 번째로는 2 바이트(한글 완성형 코드) 영문자, 숫자 및 그리스 문자의 사용으로 인한 검색율의 저하이며 네 번째로는 일부 화합물명의 오기(오타)이다. 이러한 사례는 영어권 문헌에서 OCR로 입력할 때 자주 발생하는 오류로 보고된 바 있으나 국내 특허 정보의 경우에는 입력과정에서 잘 못에 의해 발생한 것으로 보인다. 예를 들면 'metyl'가 'methyl'과 'methyi'로 입력되어 있다. 이외에도 번역오류 등을 찾아 볼 수 있다.

4) 한글맞춤법 제50항에 "전문 용어는 단어별로 띄어 씀을 원칙으로 하되, 붙여 쓸 수 있다. 전문 용어란, 특정의 학술 용어나 기술 용어를 말하는데, 대개 둘 이상의 단어가 결합하여 하나의 의미 단위에 대응하는 말, 곧 합성어의 성격으로 되어 있다. 따라서 붙여 쓸 만한 것이지만, 그 의미 파악이 쉽도록 하기 위하여 띄어 쓰는 것을 원칙으로 하고, 편의상 붙여 쓸 수 있도록 하였다. "라고 정의하고 있다.

표 2는 발명의 명칭에 나타나는 표기상의 문제점 몇 가지를 보여주고 있다.

이러한 모든 문제는 단순히 검색만 수행한다면 검색 시스템의 성능에 따라 달라질 수 있다. 실제 "초자연음이온가려움증제거비듬제거탈모방지발모촉진제"를 검색화면에 입력하면 동일한 데이터가 검색된다는 사실이다. 물론 주요 키워드 몇 개를 골라 검색어로 사용하면 경우에 따라서는 정확히 검색이 되지 않거나 불필요한 정보가 검색된다. 정보의 정량적 분석을 위해서는 검색 시스템에서 제공하는 다양한 언어학적 기법에 의존하기보다는 단어(띄어쓰기로 구분되는) 수준의 통일성이 필요하며 검색시스템의 기능에 따르면 단어가 달리 사용되거나 오타 등의 오류 데이터는 분석 대상에서 제외되기 쉽다. 발명의 명칭에 대해 445건을 무작위 추출하여 분석해 본 결과 대부분 띄어쓰기와 외래어 표기법의 문제이지만 17.5%에 달하는 78건에서 문제점을 발견할 수 있었다.

Ⅲ. 개선방안 및 결론

서론에서 기술한 바와 같이 특허정보는 매우 유용한 정보이며 기술정보인 동시에 다양한 측면에서 활용 가능한 가치를 지니고 있음에 불구하고 우리나라 특허정보의 경우는 단순히 표기법의 혼란으로 인해 활용 가치가 떨어질 뿐만 아니라 경우에 따라서는 사용자가 하나하나 수정해야만 쓸모가 있는 정보가 된다. 이를 위해서는 일차적으로 특허를 출원하는 출원자나 이를 대리하는 대리인이 한글 맞춤법을 정확히 지키는 일이 중요하다. 그러나 많은 출원자로 하여금 자발적으로 맞춤법을 지키도록 일임하는 데에는 현재와 같은 문제가 여전히 남아있게 된다. 따라서 주소의 기록단위(예, '서울', '서울특별시', '동경', '도쿄도' 등)에 대한 규정을 명확히 하는 한편 그림 1과 같은 검정 시스템을 통해 일정 기준(예를 들어 0.1%)이하의 오류가 발생하는 경우에 한하여 특허 출원을 받는 등의 조치가 필요하다. 허용기준은 점진적으로 강화시켜 궁극적으로는 오류가 없는 고품질의 특허정보가 생성되도록 개선해 나가야 할 것이다.

특허는 국가 기관이 일정한 배타적 권리를 부여하는 문서이다. 자기 나라에서 정한 어문의 맞춤법을 지키지 않는 문서에 권리를 부여해 주는 문제가 있다. 단순히 특허문서에 수록된 정보의 정확한 유통 문제만이 아니라 국가 언어문화의 수준 문제이다. 특허정보가 유용한 정보로서 보다 큰 가치를 발휘할 수 있도록 하기 위해서는 특허정보의 중요성과 활용 가치에 대한 인식의 확산과 더불어 활용하기 쉽도록 정확한 데이터를 위한 제도적, 기술적 장치가 필요하다.

■ 참 고 문 헌 ■

[1] "IP and business: Patent Information: Buried Treasure", WIPO Magazine, Jan., 2005

[2] 남영준, 정의섭, "인용정보를 이용한 신 특허지수 개발에 관한 연구", 정보관리학회지, 23(1), pp.221-241, 2006

[3] 박현우, 김기일, "특허정보를 통한 PMP 연구동향과 기술경쟁력 분석", 한국콘텐츠학회논문지, 7(9), pp.117-126, 2007

[4] Thielemann, W., "OCR errors in patent full-text documents", IRF symposium 2007, 8-9 Nov., Vienna, http://www.ir-facility.org/symposium/irf-symposium-2007/videos-and-presentations, 2007

[5] 조순영, "종합목록 데이터의 오류 유형에 관한 연구 - KERIS 종합목록의 학위논문 서지데이터를 중심으로", 한국문헌정보학회지, 36(4), pp.5-19, 2002

[6] 윤정옥, "연속간행물 종합목록 데이터베이스의 레코드 품질 평가", 한국문헌정보학회지, 37(1), pp.27-42, 2003

[7] Onwuebuzie, A.J., Waytowich, V.L., and Jiao, Q.G., "Bibliographic errors in articles submitted to scholary journals: the case for research in the schools", http://asstudents.unco.edu/students/AE-Extra/2006/12/Jiao.html

[8] Pandit, I., "Citation errors in library literature: a study of five library science journals", Library and information science research, 15(2), 185-198, 1993

[9] Lok, C.K., Chan, M.T., and Martinson, I.M., "Risk factors for citation errors in peer-reviewed nursing journals", Journal of advanced nursing, 34(2), pp.223-229, 2001

[10] 한국데이터베이스진흥센터, "데이터베이스 품질평가에 관한 연구", 한국데이터베이스진흥센터, 1995

[11] 한국데이터베이스진흥센터, "데이터베이스 품질평가 항목", 한국데이터베이스진흥센터, 2000

[12] 이응봉, 조현양, 류범종, 최재황, "과학기술분야 데이터베이스의 품질향상을 위한 품질평가 연구", 한국문헌정보학회지, 35(2), pp.109-132, 2001

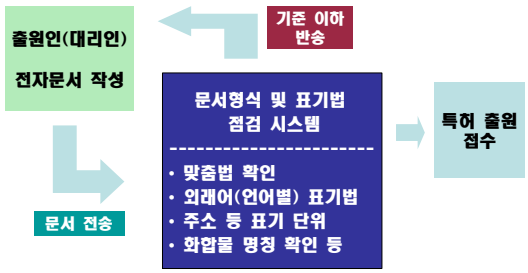
[13] 이제환, "공동목록 DB의 품질평가와 품질관리: KERIS의 종합목록 DB를 중심으로", 한국문헌정보학회지, 36(1), pp.61-89, 2002

표 1. 국가명 및 주소 표기 사례

구분	사례				
국가명	미국 미합중국 ---- 일본 일본국	독일 독일국 독일연방공화국 ---- 프랑스 프랑스공화국	스위스 스위스 체하 스위스 씨에이치 스위스 시에이취 스위스연방 스위스연방국 스위스연방공화국 스위스연방 체하 스위스연방 씨에이치	싱가폴 싱가포르 --- 터키 터어키 --- 말레이시아 말레이지아	네델란드 네덜란드 네델란드왕국 네덜란드왕국 ----- 대만 중화민국 중화민국 공화국
도시/주명	동경 도쿄 도꾜 토쿄 동경도 도쿄도 도꾜도 토쿄도 토꾜토 토쿄토 도오꾜도 도오쿄오도 토오쿄오토 도오꾜오도	뉴저지 뉴저지주 ----- 치요다구 치요타쿠 치요다꾸 ----- 고마츠시 고마츠-시 ----- 요코하마 요코하마시 요코하마-시	노스.... 노쓰.... 노오스....	웰링턴 웰링톤 ---- 워싱턴 워싱톤 ---- 멜버른 멜보른	산호세 산조세 산조스 센호세
기업명	퀼컴 콸콤 니뽄 니혼 니폰 닛본	코오포레이션 코오포레이숀 코오퍼레이션 코포레이션 코퍼레이션 코포레이트 코프.	인크 잉크 인크. 인코오퍼레이티드 인코포에이티드 인코퍼레이티드	캄파니 캄퍼니 컴파니 컴퍼니	가브시키가이샤 .... 가브시키 카이샤 ....내 ....나이
기타	죠메 죠오메 초메 쵸오메	스트라세 스트라쎄 슈트라세	데이빗 데이비드 스프링스 스프링즈	에너지 에네루기	피오박스 피.오.바스 피오.박스 피오 박스 ....

표 2. 발명의 명칭에 나타나는 표기 문제의 사례

구분	사례		
띄어쓰기	발바닥기능회복교정깔창 시스템에서SETMANAGEMENTSAMECHANNELPARAMETERS 양방향원격제어통합그래픽유저인터페이스운영방법 초자연음이온가려움증제거비듬제거탈모방지발모촉진제 할로알킬알콕시실란으로부터오르가노알콕시실란을 형상기억합금을이용한열원차동력발생장치 빛반도체		
외래어 표기	플라즈마 프라즈마	가스켓 가스킷 개스킷	콘텐트 컨텐츠 콘텐츠
2바이트 문자	△Σ-형 Ⅲ-Ⅴ족 κ 3-BH3을 γ-오리자놀 αβ-불포화	2상 c—k i t ZSM-48 V i d e o B r o a d c a s t i n g e I F-5A1s i RNA	
오타 등	methyi 신규한 중간체(NOVEL INTERMEDIATES) 공막돌륨슬 밴드 내버게이션 서비스(NAVIGATION SERVICES) 노-쥬브 깔대기/ 깔때기 낚시대/ 낚싯대 다이나믹/ 다이내믹 네스트된/ 네스티드/ 네스팅된		



▶▶ 그림 1. 개선 시스템 개념도