특허정보 전문검색을 위한 문헌구조화 연구

A Study on Patent Structure in Patent Full-text Retrieval

권영숙, 이두영
중앙대학교 문헌정보학과

kwon, young-sook, Lee doo-young.
Dept. of Library & Information Science, Chung-Ang Univ.

특허정보는 일반 과학기술정보와 다른 특성을 가지고 있어 정확성과 최신성이 절대적으로 필요하다. 이와 같은 특허정보의 특성을 고려하여 이용자의 정보요구를 충족시키고 효과적으로 검색할 수 있는 특허정보검색시스템 구축을 위한 기초자료로서 특허문헌구조를 고찰하였다.

1. 서론

오늘날 기술개발의 성패여부는 기술의 변화를 미리 예측하고 이에 어떻게 대처하느냐에 달려있다. 기술변화를 확실하고 신속하게 접할 수 있는 가장 좋은 자료는 바로 문헌적인 성격과 권리적인 성격을 겸비한 특허정보이다. 특허정보는 일반 과학기술정보와 상이한 특징을 지니며, 일반 과학기술정보를 능가하는 이점을 가지고 있다. 또한 특허정보는 창조적인 기술개발에 의하여 산업의 지식집약화를 가능하게 하는 필수적인 요소로 인정되고 있다.

특히, 특허정보는 정보의 속성상 정확성과 최신성이 절대적으로 필요하며 정보이용자는 특허에 대한 기술정보 뿐만 아니라 권리정보도 함께 탐색해야하는 이중의 부담이 있으며, 특히 권리정보에 대한 사항은 법률지식이 필요로 되기 때문에 탐색기법에 대한 지식이 부족한 이용자들도 특허정보를 쉽게 검색하고 이해할 수 있는 시스템상의 장치가 필요하다.

한편 특허출원 건수는 매년 증가하고 있어 현재 전세계적으로 년간 120만건 정도의 특허정보가 발생되고 국내에서는 년간 10만건 정도가 발생하고 있다. 특허정보검색은 각 국가에서 출원된 정보 모두가 빠짐없이 수록된 정보원에서 탐색해야 하기 때문에 이용자요구에 적합하지 않은 특허정보의 검색은 억제하고, 적합정보만을 검색해 내야 하는 부담이 따른다.

이상과 같은 특허정보의 특성을 고려하여 이용자의 정보요구를 충족시키고 효과적으로 검색할 수 있는 특허정보검색시스템이 필요하게 되었다. 현재 운영하고 있는 국내외 특허정보 데이터베이스가 이용자의 정보요구를 만족시키고 효과적인 검색을 하기에는 여러 가지 문제점이 있다.

따라서 본 고에서는 특허정보 이용자들의 정보요구에 부응하기 위하여 실제 이용자들의 요구를 수렴한 특허정보 전문검색시스템 구축을 위한 기초자료로서 특허문헌 내용의 구조화에 관하여 고찰해 보고자 한다. 실제 시스템의 구현과 이에 대한 평가는 차후로 미루고자 한다.

2. 특허정보의 특성 및 특허정보 관리

2.1 특허정보의 특성

특허정보는 권리정보의 면에 있어서 일반문헌과 비교하여 전혀 다른 특성을 가지고 있지만 기술정보로서도 일반문헌에 비하여 다른 특성을 갖고 있다. 예를 들면, 일정 기술수준을 초월하고 있다거나, 비교적 상세하게 그 기술내용이 기재되고 있고, 전 기술분야에 대하여 수록되고 분류가 부여되어 조사가 편리하며 특정 기술에 대하여 문헌발행량, 축적량 등을 정확히 파악할 수 있고, 외국인의 발명도 국어로 읽어볼 수 있으며 기술의 배경, 문제점, 미해결점 등을 알 수 있고, 문헌의 입수가 용이하며 기업의 기술개발 동향을 파악할 수 있다는 특성이 있다.

2.2 특허정보 관리

특허정보 관리의 필요성에 대한 주요한 이유로서 방대한 특허정보의 축적량과 그 현저한 증가 경향을 들 수 있다.

정보의 입수와 활용에 큰 비중을 두고 있는 현대의 과학사회에서는 세계에서 공개되는 기술적 발명에 대하여 항상 초점을 두고 확실하

게 파악하고 있어야 한다. 그런데 이와 같이 특허정보를 과학자 및 기술자들이 수집. 정리하고 최신정보 조사와 소급정보를 통해 연구에 관련 있는 정보를 전부 탐색해 내는 데는 엄청난 시간과 노력을 요구하는 것이다.

또한 특허정보는 권리정보라는 특수성 때문에 관련정보 중에서 한 건만 누락되어도 연구의 중복 및 경제적 손실을 가져오는 경우가 있으므로 정보탐색에서 상당히 높은 적합률이 요구된다.

3. 특허정보검색

온라인이나 웹 등을 통해서 제공되는 전문정보 중에서 특허정보 만큼 조직화가 잘 되어 있는 것도 드물다. 특정 기술을 발명하여 그 권리를 받기 위해서는 필수적으로 수백만 건의 선행 특허정보를 조사해 보아야 한다.

또한 특허정보는 데이터베이스의 형태로 조직화가 잘 되어 있는 반면, 일반인이 쉽게 접근할 수 없는 전문정보이다. 초보자로서는 어디에 어느 정보가 있는지를 아는 것도 어려우며 알았다 할지라도, 각각 데이터베이스의 내용과 특징을 알아 정작 필요한 내용을 검색하는 것은 매우 어려운 일이다.

특허정보의 검색시 번호조회와 같이 특허정보를 출원번호나 분류 등의 서지사항으로 검색이 가능한 것과 권리나 기술에 관한 주제검색의 경우와 같이 보다 복잡한 검색을 필요로 하고 전문적인 특허에 대한 지식을 필요로 하는 경우가 있다. 또한 특정 기술에 대한 최신동향 조사는 온라인을 통하여 방대한 자료를 입수하는 것보다 관련 특허CD-ROM을 통하여 조사하는 것이 필요한 경우도 있다.

특허정보 검색에 있어서 가장 중요한 것은 각 데이터베이스 내에서 초록 수록 유무와 전문(Full-text)의 제공여부를 확인하는 것이다. 또한 특허정보 데이터베이스가 어떤 항목들을 수록하고 있는지를 파악하여 검색 가능한 것과 불가능한 것을 판단해야 한다.

특허정보 검색시 중요한 검색항목으로는 서지사항, 청구범위, 초록, 도면, 전문명세서 등이라고 할 수 있다. 산업기술정보원에서 제공하는 데이터베이스의 경우 특허, 실용신안 공개, 공고에서 청구범위가 일부 제공되고 있으며 특허기술정보센터의 KIPRIS는 초록과 대표도면, 전문명세서를 제공하고 있다. 그러나 실제 검색시 데이터가 존재하지 않은 경우가 많고, 데이터베이스 구축방법이 이미지기반시스템으로 전문에 나타나는 문자들이 페이지 단위의 이미지로 디지털화되기 때문에 컴퓨터에 의한 전문검색 및 내용 파악이 불가능하며, 단순히 전문가가 선정한 키워드에 의해서 탐색이 결정된다는 제한점이 있다. 이러한 이미지기반 전문데이터베이스의 문제점을 해결하기 위한 방법으로 문헌내용의 구조화를 통한 특허정보 전문검색시스템구축이 이루어져야 할 것이다..

4. 특허문헌 내용의 구조화

ISO 8879의 태그집합은 특허문헌에 대한 태그로 정의되지 못하기 때문에 유럽 특허청에서는 특허문헌 처리를 위한 SGML태그를 개발하였는데 이것은 특허문헌 텍스트의 공통부호화를 위한 표준포맷(ST.32)이다. DTD 및 SGML 태그의 개발은 문헌의 내용, 데이터요소 등을 이해하여 문헌이 마크업될 수 있도록 문헌의 분석이 요구된다. 다음에서 특허문헌을 위하여 개발된 마크업을 살펴본다.

4.1 특허문헌의 구조 및 마크업

ST.32에 이용된 SGML태그의 계층은 특허문헌의 일반구조를 다룬다. 계층의 단계는 일반적이며 논리적인 문헌요소를 기술하는 SGML 태그에 의하여 지시되고 문헌요소는 전체문헌, 특정 서브문헌, 문단, 표 등과 같은 텍스트의 구성요소이다. 이들 문헌요소의 각각에 대한 시작태그와 종료태그는 아래와 같이 기술될 수 있다.

Level tag(example)	SGML
Document	<PATDOC>
Sub-document	<SDOBI>
Text Component(Paragraph)	
Text Element(Subscript)	<SB>
Character	
End	</SB>
End	
End	</SDOBI>
End	</PATDOC>

특허정보는 기술정보로서 특허정보와 권리정보로서 특허정보가 있는데 기술정보로서의 특허정보는 대체로 6개 부분으로 구성되어 있다.

(1) 서지데이터: 출원자, 출원자주소, 발명자, 발명의 명칭, 특허분류번호, 파

일링날짜, 출원번호 등 이러한 데이터는 표제지에 수록되며 이들 데이터의 대부분은 출원자에 의하여 등록된다.

(2) 초록: 발명에 대한 간략한 기술. 이것은 일반적으로 표제지에 인쇄되며 그림초록이 포함되기도 한다. 초록의 분량은 보통 1페이지 이하이다.

(3) 발명의 명세서: 발명에 대한 상세한 설명으로 가장 긴 문서이다. 국내 특허출원의 경우 대략 10-20페이지 정도이나 100페이지 이상이 될 수도 있다.

(4) 청구범위: 법적보호를 요구하는 것으로 1-1 이상의 페이지로 구성된다.

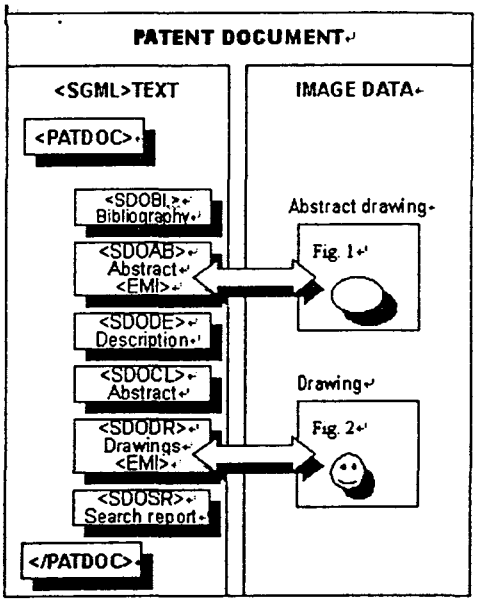
(5) 도면: 일반적으로 별도의 문서로서 기술도면, 흐름도, 다이어그램, 그래프 등을 포함한다.

(6) 탐색보고서: 탐색결과를 제시하는 것으로 보통 1페이지 정도이며 특허청에 의하여 제공된다.

이러한 부분은 하위 특허문서로서 각 문서마다 ST.32에 의하여 다음과 같이 표시된다.

<PATDOC>
Start of patent document
  <SDOBI>
    Bibliographic data
  </SDOBI>
  <SDOAB>
    Abstract
  </SDOAB>
  <SDODE>
    Description
  </SDODE>
  <SDOCL>
    Claims
  </SDOCL>
  <SDODR>
    Drawings
  </SDODR>
  <SDOSR>
    Search report
  </SDOSR>
  End of patent document
</PATDOC>

외부 엔티티인 이미지데이터와 연결되어 있는 이러한 구조는 그림1에서 설명하고 있다.



<그림 1 >

4.1.1 서지 태그

서지 태그에 대한 구조는 WIPO Standard ST.9와 ST.30을 기초로 하고 있다. ST.9 표준에 있는 코드는 특허문헌의 표제지에 프린트되는 주요 데이터요소를 위해서 제공되는데, 이것을 INID코드라고 한다. 서지데이터에 대한 DTD설계에서 SGML태그의 기본인 ST.9와 ST.30표준코드를 사용한다. 이 데이터의 계층구조는 다음과 같다.

<SDOBI> Start of bibliographic data
  <B100> INID code 10 - identification of the patent
  <B200> INID code 20 - application data
  <B300> INID code 30 - priority data
  <B400> INID code 40 - public availability
  <B500> INID code 50 - technical information
  <B600> INID code 60 - related documents
  <B700> INID code 70 - parties concerned with the patent
  <B800> INID code 80 - data related international conventions
</SDOBI> End of bibliographic data

서지데이터 및 태그의 경우에 특허출원자가 SGML을 사용한다면 서지데이터 태그에 대하여 전혀 걱정할 필요가 없는데 그것은 데이터

엔트리가 소프트웨어에 의하여 처리되기 때문이다.

4.1.2 초록, 명세서, 특허청구범위

이들 서브문서는 일반적으로 출원자에 의하여 등록되는 데이터인데 학술문헌과 유사하다. 이런 점에서 데이터 마크업에 사용된 태그가 특허문헌에만 필요한 것이 아니므로 이 부분에 대한 태그를 다시 설계하지 않고 학술문헌에 사용하는 태그가 특허문헌에 적용될 수 있는지를 결정하여 사용한다. 그러나 일반적으로 일반문헌에 사용하는 태그는 특허문헌에 맞지 않는다. 예를 들어 내용 중에 있는 표, 색인, 장, 상호참조 등의 부분에서는 적용이 불가능하다.

4.1.3 텍스트(General text)

표목, 문단, 표 등과 같은 일반적 텍스트를 위하여 ISO 8879는 ISO/IEC/TR9573과 마찬가지로 몇가지 지침을 제시한다. <H> 표목; <P> 문단; <OL> 순차표; <LI> 표항목과 같은 태그는 ST.32와는 다른 SGML응용이며 DTD이다. 특허문헌은 일반문헌과는 달리 분모, 분자 같은 수학적인 문자가 상당히 일반적으로 나타난다. 그러나 ISO 8879는 수학적인 특수문자에 대한 마크업으로 맞지 않으므로 수식에 대한 마크업은 별도의 지침을 따른다.

4.1.4 표 (Tables)

특허문헌에서 표의 삽입은 흔히 나타나는데, 표는 마크업의 가장 중요한 부분을 차지한다. 표는 특정 소프트웨어 "Wordperpect"를 사용하여 태깅한다. ST.32는 표 마크업에 맞지 않으므로 개선할 필요가 있으며 어떤 표는 이미지 데이터로 처리되기도 한다.

4.1.5 수식(Mathematics)

표와 마찬가지로 수식도 특허문헌에서 아주 일반적으로 나타나며 마크업도 복잡하다. 수식에 대한 ST.32 DTD는 ISO 기술보고서 9573의 DTD를 유사하게 따르며 다양한 수학적 구조를 코드화 할 수 있다.

4.1.6 도면과 이미지데이터

대부분의 특허문헌은 텍스트데이터에 첨부된 별도의 문서로서 도면을 포함하고 있다. 그러나 이미지데이터(화학식, 특수문자, 복잡한 표 등)는 초록, 명세서, 청구범위와 같은 본문 텍스트 안에 나타날 수 있다. ST.32는 별도 페이지의 도면까지 포함하여 모든 이미지를 삽입이미지(tags<EMI>)로 분류한다. 이미지는 본문 텍스트외부의 이미지파일로서 외부 엔티티 처리를 하는데(그림1참조), 외부파일에 이미지가 저장되는 방법은 특허문헌의 특정요구에 따라서 변할 수 있으므로 ST.32와 다른 표준인 WIPO표준 ST.33을 사용하여야 한다.

특허 본문의 초록, 명세서, 청구범위 내에 이미지가 나타나는 경우가 있는데, 그것 또한 삽입이미지로 처리된다.

4.2 탐색 보고서

탐색보고서의 데이터는 특허청에서 등록하는 서지데이터와 유사하고 태그는 특허문헌에 대한 세부사항이다.

5. 결론 및 향후과제

정보환경의 변화에 따라 전통적인 인쇄매체에 수록되는 특허문헌의 내용이 점차 컴퓨터를 이용한 전자문헌의 형태로 전환되고 있다. 최신의 컴퓨터시스템과 다양한 온라인 통신망을 이용하여 산업계, 변리사사무소, 연구소, 학계, 발명가 등 산업재산권정보를 필요로 하는 모든 이용자의 요구에 부응하는 적합정보를 검색할 수 있는 효율적인 특허정보 전문검색시스템을 구축하기 위한 기초연구로서 특허문헌구조화에 관하여 고찰하였다. 앞으로 특허문헌 요소의 특성 및 이용자 요구분석을 반영한 실제 시스템의 구현과 이에 대한 성능평가가 수반되어야 할 것이다.

참고문헌

방용조, 특허정보 검색효율 증대 방안에 관한 연구, 연세대 석사, 1988.

신현호, 특허정보 검색시스템에 관한 연구, 건국대 경영대학원 석사, 1984.

이영주, 우리나라에서의 특허정보활동에 관한 연구, 연세대 석사, 1984.

장태종, 종합특허정보해설, 산업기술정보원,1997

Goldfarb. C. F., The SGML Handbook: The Annotated Full Text of ISO 8879 - Standard Generalized Markup Language, Oxford University Press. 1991.

Paul Brewin, "SGML and Patent Document Processing", World Paten Information 18(4), 183-192, 1996.