국제특허분류 클러스터링을 이용한 특허 검색 시스템

Patent Search System Using IPC Clustering

김한기, 이석형, 윤화묵
한국과학기술정보연구원

Kim Han-Gi, Seok-Hyoung Lee, Yoon Hwa-Mook
Korea Institute of Science and Technology Information

요약

지적재산권의 중요성이 커지면서 특허 검색을 이용하는 일반 사용자의 숫자가 늘어나고 있다. 일반적으로 한 두 개의 키워드만을 사용하는 일반 사용자의 검색 패턴을 고려할 때, 대량의 특허 문서에서 원하는 검색 결과를 찾는 일은 쉽지 않은 일이다. 이에 모든 특허 문서에서 제공되는 국제 특허 분류(IPC) 정보를 사용해서 사용자의 검색 결과를 클리스터링하여 보여주어 사용자가 검색하고자 하는 검색범위를 손쉽게 제한 할 수 있도록 도와주어 원하는 결과를 좀 더 빠르게 찾을 수 있는 특허 검색 시스템을 소개하고자 한다.

Abstract

The importance of intellectual property right becomes larger and the number of the person who uses a patent search is increasing. When considering the search pattern of the general user who uses only one or two search terms, it is not easy task to find desirable search result in the massive patent documents. So we present patent search system based on IPC Clustering which helps users confine the search result by using international patent classification (IPC) which provided from all patent documents. By using this system, the general users can find patent search result more effectively.

1. 서 론

산업화 사회에서 지식정보화 사회로 시대가 변해가면서 특허, 실용실안 등과 같은 지적재산권의 중요성이 더욱 커져가고 있습니다. 급격한 기술의 발달로 인해 해마다 출원되는 특허 출원건수는 날로 증가하고 있으며, 이제는 컴퓨터의 도움이 없이 원하는 특허 문서를 검색하는 일은 거의 불가능할 정도로 전체 누적 건수도 많아졌습니다.

특허에 대한 사회의 관심이 높아지면서 특허 분야 전문가가 아닌 일반 사용자의 특허 검색이 늘어나고 있으나, 일반적인 웹 포탈의 검색 방식[2](한 두 단어의 키워드만을 사용하고, 특허 검색 시 일반적으로 사용하는 불리언 검색 방식이 아닌 벡터 검색 방식)에 익숙한 일반 사용자가 대용량의 특허 데이터에서 원하는 검색 결과를 찾는 일은 쉽지 않은 일이다. 특히 특허 문헌은 논문이나 그 밖의 다른 학술 문헌과는 달리 국가기관에 의해서 관리되며, 통일된 고유한 서지 정보를 가진다. 특허문헌의 중요한 서지 정보 중의 하나로 국제 특허 분류 필드가 있다. 이 필드 값은 해당 특허가 속하는 기술 분야에 대한 정보를 포함하고 있으므로 이 정보를 활용하면 전체 특허 문헌에서 사용자가 원하는 특허를 제한하여 찾는 데 많은 도움을 줄 수 있다. 본 논문에서 일차적으로 사용자의 검색을 입력 받고, 해당 검색 결과를 국제 특허 분류 필드 값을 이용하여 클러스터링한 결과를 사용자에게 함께 제시하여 사용자가 원하는 특허를 좀 더 편하게 찾을 수 있는 특허 검색 시스템에 대해서 소개하고자 한다.

2. 국제특허분류

국제 특허 분류(International Patent Classification, 이하 IPC)는 1975년 발효된「특허분류에 관한 스트라스부르그(Strasbourg) 협정」에서 채택된 국제적으로 통일된 특허분류기준으로써, 전체 기술 분야를 섹션, 클래스, 서브클래스, 그룹의 4개의 레벨로 세분화하여 표현하는 계층적 시스템이다. 각각의 특허 문서에는 적절한 IPC가 할당되며, IPC의 할당은 특허 문서를 출원하는 기관에서 담당하고 있다. IPC는 특허 문서에서 선행 기술을 검색하기 위해서 없어서는 안 될 요소이다. 국제 특허 분류 체계는 새로 개발된 기술 분야를 추가하거나 시스템을 향상시키기 위해서 정기적으로 개정되고 있으며, 현재 사용되고 있는 IPC 버전은 2006년 1월 1월에 발효된 8판이다.[5]

2.1 국제특허분류의 구조

IPC 는 계층적 시스템으로 섹션(section), 클래스(class), 서브클래스(subclass), 그리고 약 7만 개의 그룹으로 구성되어 있다. (대략 10%가 메인 그룹이고 나머지는 서브 그룹)

2.1.1 섹션 (Section)

각 섹션은 하나의 알파벳 대문자로 표시된다.

A: 생활필수품 (Human Necessities)
B: 처리조작 (Performing Operations, Transporting)
C: 화학, 야금 (Chemistry, Metallurgy)
D: 섬유, 종이 (Textiles, Paper)
E: 고정구조물 (Fixed Constructions)
F: 기계공학, 조명, 가열, 무기 (Mechanical Engineering, Lighting, Heating, Weapons)
G: 물리학 (Physics)
H: 전기 (Electricity)

2.1.2 클래스 (Class)

각 클래스는 관련된 섹션 이름 다음에 나오는 두 자리 숫자로 이루어진다. 예를 들어 서브섹션 "식료품;담배"의 경우 다음의 네 개의 클래스로 이루어져 있다.

A21 제빵 ; 반죽 제조 또는 가공의 기계 혹은 설비 ; 제빵용 반죽
A22 도살; 육(肉) 처리; 가금류 또는 어류의 가공
A23 다른 클래스에 속하지 않는 그것들의 처리; 식품 또는 식료품
A24 담배(TBACCO); 엽권담배(CIGARS); 지권담배(CIGARETTES); 흡연용구

2.1.3 서브클래스(Subclass)

각 서브클래스는 관련된 클래스 이름 다음에 나오는 알파벳 대문자로 표시된다. 예를 들어 클래스 A21 의 경우, 다음과 같이 3개의 서브클래스로 나누어진다.

A21B 제빵용 오븐(Oven); 제빵용 기계 또는 장치
A21C 가루반죽 제조와 가공용 기계 및 설비; 가루반죽으로 제조된 빵류의 취급
A21D 제빵 앞 또는 중간의 첨가물에 의한 곡분 또는 반죽의 처리

2.1.4 그룹(Group)

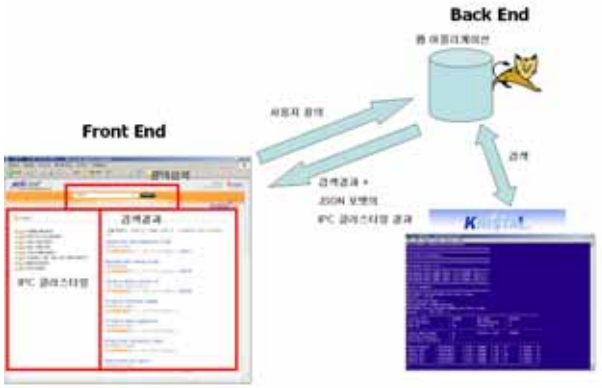
각 그룹은 서브 클래스 다음에 위치하며 슬래시(oblique stroke)로 구분된 두 개의 숫자로 구성된다. 슬래시 앞 쪽에 있는 첫 번째 숫자는 한 두 자리, 혹은 세 자리 숫자도 될 수 있다. 두 번째 숫자는 두 자리부터 다섯 자리까지 숫자로 이루어진다. 메인 그룹의 경우 두 번째 숫자는 두 자리의 '00' 으로 표시된다. 예를 들어 서브 클래스 A21B의 경우 5개의 메인 그룹(1/00, 2/00, 3/00, 5/00, 7/00)을 가지며, 이 중 처음 두 개는 다음과 같다.

A21B 1/00 제빵용 오븐
A21B 2/00 고주파 또는 적외선의 가열에 의한 제빵 장치
메인 그룹 A21B 1/00 는 19 개의 서브 그룹으로 나누어지며, 이 중 처음 4개는 다음과 같다.
A21B 1/02 – 가열배치에 특징이 있는 것
A21B 1/04 –– 빵을 굽기 전에만 불로 가열하는 오븐
A21B 1/06 –– 라디에이터(radiator)에 의한 것
A21B 1/08 ––– 증기가열의 라디에이터에 의한 것

위의 예에서 볼 수 있는 것처럼, 모든 서브 그룹이 같은 계층 구조상에 있는 건 아니다. 이 중 가장 상위에 있는 건 대시(dash)가 하나 있는 것이며, 그 다음으로 대시의 숫자가 많아질수록 하위 계층에 속한다.

3. 시스템 구성

전체 시스템은 리눅스 시스템에서 아파치 톰캣(Apache Tomcat) 웹 컨테이너[6]를 사용해서 구축되었으며, 검색 결과의 IPC 클러스터링 서비스를 제공하기 위해서 KRISTAL-IRMS[7]의 group by 기능을 사용하였다. 시스템은 크게 사용자의 입력을 받고, 검색 결과를 보여주는 Front End 부분과 사용자 질의를 받아서 특허 문서를 검색하고, 해당 검색 결과의 IPC 클러스터링을 구하는 Back End 부분으로 나뉜다.



▶▶ 그림 1. 전체 시스템 구성도

3.1 Front End

Frond End 부분은 사용자의 질의를 입력 받는 부분, 사용자

의 질의에 대한 검색 결과를 보여주는 부분, 그리고 검색 결과에 대한 IPC 클러스터링을 보여주는 3부분으로 이루어져 있다. 각 부분은 JSP(JavaServer Pages)와 Javascript를 이용한 웹 애플리케이션으로 구현되어 있으며, IPC 클러스터링을 보여주는 부분과 검색 결과를 보여주는 부분은 Ajax (Asynchronous JavaScript and XML) 기술을 적용하여, 사용자 질의가 입력되면 각각 Back End 부분과 따로 통신하여 해당 결과를 사용자에게 보여준다.

Frond End와 Back End사이의 데이터 교환은 JSON (JavaScript Object Notation)[9] 방식을 사용하여, 둘 사이에 이동되는 데이터의 크기를 최소화 하였으며, JSON 데이터 처리에는 Prototype JavaScript Framework[8]를 사용하였다. IPC 클러스터링 결과에 대한 JSON 방식의 데이터 구조는 표.2에 나와 있으며, 검색어 "자동차"에 대한 IPC 클러스터링 결과는 다음과 같다.

[표 1] 검색어"자동차"에 대한 IPC 클러스터링 결과(JSON 포맷)

{"results":[{"level":"1","count":692,"name":"A","desc":"A 생활필수품"},
{"level":"1","count":66228,"name":"B","desc":"B 처리조작; 운수"},
{"level":"1","count":746,"name":"C","desc":"C 화학; 야금"},
{"level":"1","count":72,"name":"D","desc":"D 섬유; 지류"},
{"level":"1","count":2938,"name":"E","desc":"E 고정구조물"},
{"level":"1","count":14041,"name":"F","desc":"F 기계공학; 조명; 가열; 무기; 폭파"},
{"level":"1","count":3043,"name":"G","desc":"G 물리학"},
{"level":"1","count":3308,"name":"H","desc":"H 전기"}]}

[표 2] IPC 클러스터링 결과에 대한 JSON 구조체

문자열	값 (value)
level	1부터 4 사이의 값을 가진다. 1:섹션, 2:클래스, 3:서브클래스, 4:그룹
count	검색 결과에서 해당 IPC의 개수
name	IPC 값
desc	IPC name에 대한 설명

3.2 Back End

Back End 부분은 사용자 질의를 받아서 검색 후, 검색 결과를 Front End로 보내주는 부분과 해당 검색 결과에 대한 IPC 클러스터링을 구해서 JSON 포맷으로 변환한 후 Front End로 보내주는 두 부분으로 이루어져 있다. KRISTAL Java-API를 이용한 Java 프로그램으로 작성되었으며, 클러스터링 결과를 JSON 포맷으로 변환하기 위해서 json.org[9]에서 제공하는 Java 라이브러리를 사용하였다.

검색 결과에 대한 IPC 클러스터링을 구하기 위해서 KRISTAL-IRMS에서 제공하는 그룹바이 검색기능을 이용하였다. KRISTAL-IRMS는 여러 가지 그룹바이 기능을 제공하고 있으며[7], 여기에서는 DB의 특정 Section별로 그룹화하는 기능을 이용하였다.

3.3 대상 데이터와 데이터 가공

3.3.1 대상 데이터

테스트용으로 사용된 특허 데이터는 한국특허정보원[10]에서 제공되어 한국과학기술정보연구원[11]에서 서비스되는 한국 특허/실용실안 데이터로 2007년 6월 15일까지 입수된 2,240,511 건의 이터를 대상으로 한다.

3.3.2 데이터 가공

하나의 특허에 대해 하나의 IPC가 할당되는 경우가 일반적이기는 하나, 기술 분야 간의 협력과 복잡성으로 인해 한 특허에 여러 IPC가 할당 될 수 도 있다. 예를 들어, 출원번호 10-1997-0003845 "박막 트랜지스터 기판"(삼성전자주식회사)의 경우, IPC를 확인해 보면 H01L 27/12, G02F 1/136, H01L 29/786 와 같이 3개의 IPC가 할당되었음을 확인할 수 있다. 이와 같이 하나의 특허에 대해 다중 IPC가 할당 되어 있을 경우 첫 번째 IPC 값이 해당 특허를 나타내는데 가장 중요한 IPC 라고 가정하고, 첫 번째 값만을 사용하였다.

원시 데이터의 IPC 필드 값에 들어있는 내용은 섹션, 클래스, 서브클래스, 그룹의 구별이 따로 없기 때문에 IPC 필드 값을 읽어서 신택스(syntax)를 확인한 후, 신택스에 문제가 있는 특허는 변환하지 않고, 올바른 신택스의 특허에서만 섹션, 클래스, 서브클래스, 그룹 값을 추출해서 사용하였다.

4. 결 론

이상과 같이 특허 문서에서 제공되는 국제 특허 분류(IPC) 정보를 사용하여 사용자가 원하는 검색 결과를 좀 더 편리하게 찾을 수 있도록 도와주는 특허 검색 시스템에 대하여 알아보았다. 특허 문서에서 제공되는 분류 정보를 적극적으로 활용함으로써 기존의 특허 전문가가 아닌 일반 연구자들도 보다 효율적인 특허 검색을 할 수 있을 것으로 기대된다.

▌참 고 문 헌▐

[1] 황태청, "국제특허분류(IPC)의 역사", 발명특허(kipa), 4권, pp9-11, 1979

[2] Silverstein, C. and Henzinger, M. and Marais, H. and Moricz, M., "Analysis of a very large AltaVista query log", Digital SRC, 1998

[3] Leah S. Larkey, "A patent search and classification system" Proceedings of 4th ACM Conference on Digital Libraries, pp. 179-187, 1999

[4] Kang, I.S. and Na, S.H. and Kim, J. and Lee, J.H., "Cluster-based patent retrieval", Information Processing & Management, vol43, pp. 1173-1182, 2007
[5] World Intellectual Propert Organization(WIPO) http://www.wipo.int/classifications/ipc/en/
[6] Apache Tomcat http://tomcat.apache.org/
[7] KRISTAL-IRMS http://www.kristalinfo.com/
[8] Prototype http://www.prototypejs.org/
[9] JSON http://www.json.org/
[10] 한국특허정보원 http://www.kipi.or.kr/
[11] 한국과학기술정보연구원 http://www.kisti.re.kr/