ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류☆

IPC Multi-label Classification based on Functional Characteristics of Fields in Patent Documents

임 소 라[1*] 권 용 진[1]
Sora Lim YongJin Kwon

요 약

최근 지식과 정보가 가치를 생산하는 지식기반사회로 접어들면서 지식재산권의 대표적인 형태인 특허에 대한 중요성이 매우 높아지고 있으며 출원되는 특허의 양도 매년 증가하고 있다. 방대한 양의 특허정보를 효과적으로 이용하기 위해서 특허문서를 그 발명의 기술적 주제에 따라 적절하게 분류하는 것이 필요하며 이를 위해 IPC(International Patent Classification)가 주로 사용되고 있다. 현재 주로 사람의 손으로 이뤄지는 특허문서의 IPC 분류과정의 효율성을 높이기 위하여 다양한 데이터마이닝과 기계학습 알고리즘을 기반으로 IPC 자동분류에 관한 연구들이 수행되어 왔다. 하지만 기존의 IPC 자동분류에 관한 연구의 대부분은 특허문서의 구조적 특징과 같은 특허문서 고유의 데이터 특성에 대한 고려보다는 다양한 기계학습 알고리즘을 특허문서로 적용하는 것에 초점을 맞춰왔다. 이에 본 논문에서는 IPC 자동분류를 위해 특허문서의 특징과 구조적 필드의 역할을 기반으로 특허문서 분류에 영향을 끼치는 두 가지 필드, 기술분야 및 배경기술 필드의 활용을 제안한다. 그리고 특허문서가 동시에 다수의 IPC 분류코드를 가지는 점을 반영하여 다중 레이블 분류(multi-label classification) 모델을 구축한다. 또한 IPC 다중 레이블 분류의 실제 현장에서의 적용 가능성 확인을 위해 630개의 범주를 가지는 IPC 서브클래스 레벨까지 분류 가능한 수법을 제안한다. 이를 위해 국내에서 등록된 564,793건의 특허문서를 대상으로 특허문서의 구조적 필드의 영향을 확인하기 위한 IPC 다중 레이블 분류 실험을 수행하였고, 그 결과 제목, 요약, 청구항, 기술분야 및 배경기술 필드를 활용한 실험에서 87.2%의 싱글매치 정확도를 얻었다. 이를 통해 기술분야 및 배경기술 두 필드가 IPC 서브클래스 레벨까지의 다중 레이블 분류의 정확도를 향상시키는데 중요한 역할을 하고 있음을 확인하였다.

☞ 주제어 : 특허 분류, IPC 자동분류, 특허문서 필드, 필드 기능, 멀티 레이블 분류

ABSTRACT

Recently, with the advent of knowledge based society where information and knowledge make values, patents which are the representative form of intellectual property have become important, and the number of the patents follows growing trends. Thus, it needs to classify the patents depending on the technological topic of the invention appropriately in order to use a vast amount of the patent information effectively. IPC (International Patent Classification) is widely used for this situation. Researches about IPC automatic classification have been studied using data mining and machine learning algorithms to improve current IPC classification task which categorizes patent documents by hand. However, most of the previous researches have focused on applying various existing machine learning methods to the patent documents rather than considering on the characteristics of the data or the structure of patent documents. In this paper, therefore, we propose to use two structural fields, technical field and background, considered as having impacts on the patent classification, where the two field are selected by applying of the characteristics of patent documents and the role of the structural fields. We also construct multi-label classification model to reflect what a patent document could have multiple IPCs. Furthermore, we propose a method to classify patent documents at the IPC subclass level comprised of 630 categories so that we investigate the possibility of applying the IPC multi-label classification model into the real field. The effect of structural fields of patent documents are examined using 564,793 registered patents in Korea, and 87.2% precision is obtained in the case of using title, abstract, claims, technical field and background. From this sequence, we verify that the technical field and background have an important role in improving the precision of IPC multi-label classification in IPC subclass level.

☞ keyword : Patent classification, IPC Classification, Patent Document Fields, Field function, Multi-label classification

1 Dept. of Telecommunication and Information Engineering, Korea Aerospace University, Goyang, 10540, Korea
* Corresponding author (ebbunsora@kau.ac.kr)
[Received 9 June 2016, Reviewed 14 June 2016(R2 28 September 2016), Accepted 1 December 2016]
☆ 본 논문은 2016년도 한국인터넷정보학회 춘계학술대회에서 발표한 논문을 확장한 버전 임

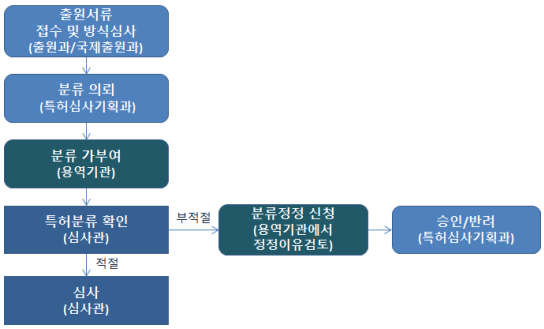
☆ This work was supported by the GRRC Program of Gyeonggi province. [GRRC2016-B01, Advanced Intelligent Ambient Broadcasting Media Content Development and Application]
☆ 온유 특허법률사무소 안병규 대표변리사의 특허 관련 귀중한 조언에 감사를 표합니다.

1. 서 론

지식과 정보가 경쟁력의 원천이 되는 지식기반 사회에 접어들면서 노동 및 자본기반의 경제활동으로 산출된 결과물에 대한 재산권보다는 지식활동으로 얻어진 결과물을 재산권 형태로 보호하는 지식재산권에 대한 중요성이 날로 커지고 있다. 이에 따라 지식재산권을 대표하는 형태 중의 하나인 특허의 출원 건수도 매년 증가하고 있으며, 2014년 한 해 동안 국내에서 출원된 특허는 약 21만건에 이른다.[1] 또한 특허의 누적 등록건수가 향후 폭발적으로 증가할 것으로 예상되는 상황에서 특허 조사를 효과적으로 수행할 수 있도록 특허를 기술 및 산업분야에 따라 적절하게 분류하는 것이 중요해 지고 있다. 특허문서의 적절한 분류는 1) 특허출원 명세서의 신규성이나 진보성을 평가하기 위하여 기존 특허문헌을 효과적으로 조사(선행기술 검색)하는데 활용되고 2) 특허문서에 포함되어 있는 기술 및 권리 정보에 용이하게 접근하기 위해 특허문서를 정연하게 배열하고 보급하는데 활용되며 3) 여러 분야에 대한 기술 발전 평가 및 통계 작성에 활용 및 4) 특허분류의 정확성 및 일관성 확보와 검색 효율성을 제고하여 특허업무의 질적 수준 향상 등을 목적으로 유용하게 사용되고 있다.[2]

아울러 특허청의 '특허·실용신안 심사기준'[3]에 따르면 특허 분류 부여는 (그림 1)과 같은 절차로 수행된다. 특허청에 접수된 일반출원이 접수 과정에서 문제가 없는 경우, 특허청은 외부용역 기관에 해당 출원의 특허분류 부여를 의뢰한다. 용역기관에서는 의뢰받은 출원들에 대해 분류원(사람)이 각 출원의 기술내용에 따라 특허분류의 분류표상의 특정 분류개소로 각 출원을 분류한다. 이렇게 용역기관에서 일반출원을 가분류한 다음에 심사관이 특허분류가 출원된 발명의 기술적 내용에 따라 적절하게 부여되었는지 여부를 확인하고 심사에 착수하는 과정을 거친다. 이 과정에서 특허 분류의 부여를 위하여 국제특허분류 (International Patent Classification, IPC)가 사용되고 있다. IPC에 대한 자세한 내용은 2장에서 기술한다. 이와 같은 인간에 의한 분류과정을 대신할 수 있는 효율적인 학습모델의 개발 및 실용적인 시스템 구축은 데이터마이닝 및 기계학습 연구의 중요한 분야 중의 하나이며 현장에서 활용가능한 실질적인 연구가 요구되고 있다.

한편, 일반적인 웹문서나 과학기술 관측 데이터와는 달리 특허 문서는 그 자체의 고유의 특징을 갖고 있으므로 단순히 진보된 다양한 기계학습 방법을 그대로 적용해서는 좋은 결과를 기대할 수 없다. 특허문서를 자동 분



(그림 1) 특허분류 부여 절차
(Figure 1) IPC labeling Process of Patents

류하고자 할 때 일반적으로 텍스트 분류에서 사용했던 방법 이외에 고려해야 할 특허문서의 몇 가지 특징들은 다음과 같다.

첫째, 특허문서는 구조화된 문서로 서지정보, 제목, 요약, 청구항, 명세서 등의 필드로 구성된다. 서지정보는 발명인, 출원일자 그리고 IPC와 같은 세부 필드로 구성되어 있으며, 명세서는 그 발명이 속한 '기술분야', '배경기술'과 더불어 '해결하려는 과제', '과제의 해결수단', '발명의 효과'를 포함하는 세부 필드로 이루어져 있다. 여기서 '기술분야' 필드는 그 발명이 속하는 산업, 기술 분야를 기술하고 있으며 '배경기술' 필드는 그 발명과 관련성이 있는 종래기술과 그 발명의 목적과 관련하여 종래의 기술이 갖고 있는 문제점을 함께 기술하고 있다. IPC는 특허문서가 속할 수 있는 기술 분야에 따라 나눠져 있으므로, IPC 분류코드로 특허문서를 분류하고자 할 때 이 두 가지 세부 필드는 특허문서 분류에 있어서 큰 영향력을 가지고 있다고 볼 수 있다.

따라서 특허문서의 주제분야를 정확하게 정의하기 시작하는 IPC 서브클래스 레벨까지의 분류에서 '기술분야', '배경기술'의 필드는 분류성능 향상에 중요한 역할을 할 것으로 사료된다.

둘째, 특허문서로서의 효력을 가지려면 신규성(novelty), 비자명성 (또는 진보성, non-obviousness), 유용성 (utility)의 세 가지 요건을 충족해야한다. 여기서 신규성은 발명의 내용이 종래의 기술에 알려져 있지 않는 새롭고 독창적인 것을 의미하며, 비자명성은 선행기술에 비해 진보한 점이 있는 것, 그리고 유용성은 산업 상에서 이용 가능한 현실적 가치를 가지는 것을 의미한다. 이 세 가지 요건은 주로 특허문서의 청구항으로 대변된다. 즉 새로운 발명에 대하여 그 아이디어의 가치를 나타내는 부분이 청구항이

다. 그러나 특허문서 자동분류에 특허문서의 새로운 발명에 대한 내용을 가장 크게 반영하고 있는 청구항 필드를 중심으로 활용하는 것은 분류 성능에 부정적인 영향을 끼칠 수 있을 것이다. 하지만 특허문서의 차별화는 청구항을 바탕으로 이뤄지므로 IPC 메인그룹 이하의 레벨과 같이 매우 세분화되어 있고 그 차이가 미미한 범주로의 특허분류에서는 청구항을 고려할 필요가 있을 것이다.

셋째, 특허문서는 일반적으로 그 특허문서에 해당하는 하나 이상의 IPC 분류가 동시에 할당되는 것이 특징이다. 이것은 기계학습이나 통계학 분야에서 분류 문제를 크게 이진 분류 (binary classification), 다중 클래스 분류 (multi-class classification), 다중 레이블 분류 (multi-label classification)로 구분할 때, 다중 레이블 분류에 해당한다. 특허문서에 특허분류를 할당하는 문제는 동시에 다수개의 IPC 분류가 할당되는 가능성을 배제한 다중 클래스 분류 문제로 접근하는 것 보다 실제 특허문서의 IPC 부여 형태를 반영한 다중 레이블 분류 방법이 실용적인 측면에서 타당할 것이다.

위와 같은 특허문서의 특징은 사람의 손으로 특허출원에 복수개의 IPC 분류를 부여하는 과정에서도 잘 활용되고 있다. 접수된 특허출원에 수작업으로 IPC 분류를 부여하는 과정에서는 특허문서의 역할과 문서의 구조에 대한 이해를 기반으로 제목, 요약과 같은 특정 필드들을 우선적으로 활용하며, 해당 특허출원의 포괄적인 기술적 범위를 특정하고 최종적으로 더욱 세세한 분류를 위해서 특허문서의 청구항을 심도있게 분석하는 것으로 알려져 있다. 따라서 특허문서 자동분류모델의 기계학습과정에서도 이와 같은 특허문서의 데이터적인 특성에 기반을 둔 사람에 의한 분류 과정을 참조할 필요가 있다고 사료된다.

그러나 IPC 기반의 특허문서 분류에 대한 기존의 연구에서는 특허문서의 고유의 구조적인 특징과 같은 데이터 그 자체의 특성에 대한 고려보다는 기계학습 모델의 선택에 집중하거나 특허문서에서 판별력있는 단어선택에 관심을 기울여 온 경향이 많다.

따라서 본 논문에서는 특허 문서의 데이터 특성에 기반을 두고 있는 사람에 의한 특허분류 과정을 충분히 고찰하여 실질적으로 활용 가능한 실용적인 IPC 자동분류를 목표로 1) 특허 문서의 각 필드 특성을 IPC 분류의 각 레벨에 따라 구별해서 활용한 기계학습 모델 구축으로 사람 손에 의한 분류과정의 대체 가능성 확인, 2) 특허 문서 분류를 현실적으로 활용 가능한 수준에서의 구축 가능성을 확인하기 위한 IPC 서브클래스 레벨에서의 기계학습 모델 구축, 3) 특허 문서 분류의 사람 개입을 최소화하기 위한 전략에서 다중 레이블 분류 가능한 특허문서 분류 모델 구축 등, 이 3가지 항목을 해결하기 위한 방법을 제안하고 약 50만 건의 특허 문서 데이터에 대해서 그 성능을 확인한다.

본 논문의 구성은 다음과 같다. 2장에서는 특허문서 IPC 분류를 수행한 기존의 연구에 대하여 설명하고, 3장에서는 본 논문에서 제안하는 특허문서 분류에 가장 영향력 있는 구조적인 필드인 '기술분야' 및 '배경기술' 필드를 사용한 IPC 자동 분류 방법에 대하여 자세히 기술한다. 그리고 4장에서는 '배경기술' 및 '기술분야' 필드를 이용한 IPC 자동 분류 실험과 결과에 대해 기술한다. 이 때 분류 모델은 가장 대표적인 머신러닝 기법인 나이브 베이즈 모델을 사용한다. 마지막으로 5장에서 결론과 향후 연구에 대하여 논의한다.

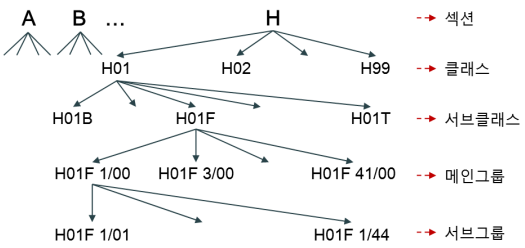
2. 관련 연구

2.1 IPC 구조

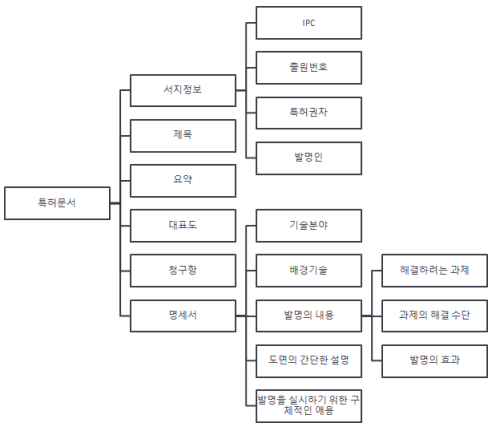
IPC는 국제특허분류(International Patent Classification)의 약자로서, 특허문서들을 고유의 분류코드로 할당한다. 이 분류코드는 특허문서에 대하여 국제적으로 통일된 분류를 하기위한 수단으로, 특허 문서가 속한 기술 분야에 따라 세분화 되어있다.

IPC 분류구조는 최상위 레벨인 8개의 섹션, 128개의 클래스, 약 650개의 서브클래스, 약 6,800개의 메인그룹, 그리고 65,000개 이상의 서브그룹의 5개의 레벨로 구성된 계층적 구조이다. (그림 2)와 같이 IPC 계층구조의 최고층 레벨인 섹션은 A에서 H까지의 8개의 알파벳으로 구분되며 A는 생활필수품, B는 처리조작, 수송, C는 화학, 야금, D는 섬유, 종이, E는 고정 구조물, F는 기계공학, 조명, 가열, 무기, 폭파, G는 물리학, 그리고 H는 전기 분야를 나타낸다. 예를 들어 'H01F 1/01'의 IPC 코드는 H 섹션 (전기), H01클래스 (기본적 전기소자), H01F 서브클래스 (자석; 인덕턴스; 변성기; 자기특성을 위한 재료의 선택), H01F 1/00 메인그룹 (자성재료를 특징으로 하는 자석 또는 자성체; 그 자성특성을 위한 재료의 선택) 내의 H01F 1/01 서브그룹 (무기재료로 된 것)에 해당한다. 이와 같이 IPC 레벨의 아래로 내려갈수록 그 기술분야는 더욱 세분화 된다.

또한 각 특허문서는 1개 이상의 IPC 분류 코드를 가질 수 있다. 예를 들면 특허 '밴디드 디스플레이를 구비한 휴대 단말기의 제어방법 및 장치 (1020140116509)*'는 G06F



(그림 2) 특허의 IPC 분류구조
(Figure 2) Structure of IPC



(그림 3) 특허문서의 구조
(Figure 3) Structure of Patents

3/048, H04B 1/40과 같이 2개의 IPC 분류 코드가 할당되어 있으며, 특허 '생명과학용 시료 보관 및 시료 관리의 통합 (1020067020977)'은 B01L 3/00, G06K 19/07, G01S 13/00, C12Q 1/68의 4개의 IPC 분류코드가 할당되어 있다. 그리고 특허 '음식물쓰레기를 이용한 고형연료 제조장치 및 제조방법 (1020150122738)'은 C10L 5/46, A47J 37/12, B01D 47/00, B01F 7/02, B02C 23/08, B04B 5/10, B07B 1/28, B09B 3/00, B30B 9/04, C10L 5/40, F26B 3/04의 11개의 IPC가 부여되어 있다.

2.2 특허문서의 구조

특허문서는 (그림 3)과 같이 크게 서지정보, 제목, 요약, 대표도, 청구항 그리고 명세서의 필드 등으로 구성되어있다. 특허문서의 첫 번째 필드인 서지정보는 IPC, 출원번호, 출원일자, 특허권자, 발명자 등의 세부 필드를 포함한다. 그 다음으로 제목, 요약, 대표도, 청구항, 명세서 필드가 따른다. 그리고 명세서 필드는 기술분야, 배경기술, 발명의 내용, 도면의 간단한 설명, 발명을 실시하기 위한 구체적인 내용과 같은 세부필드로 구성되어있다. 이와 같은 필드의 순서는 국가별로 차이가 있다. 본 논문에서는 특허문서에서 도면 필드와 같은 이미지는 제외하고 특서문서의 텍스트만을 사용하고 있다.

2.3 특허문서 분류에 관한 기존연구

최근의 빅데이터 및 데이터마이닝 분야의 기술적인 발전과 함께 다양한 기계학습 모델을 이용하여 특허 문서의 분류에 관한 연구들이 다수 수행되어 왔다. C.J. Fall 외[4]는 WIPO-alpha collection을 대상으로 Naive Bayes, KNN, SVM, 그리고 SNoW의 다양한 분류모델을 사용하여 IPC 자동 분류 연구를 수행하였다. 이들은 IPC 클래스와 서브클래스 레벨에서 특허문서 분류를 수행하였으며 특허문서에서 제목, 요약, 청구항, 그리고 처음 300개의 단어를 대상으로 IPC 자동분류를 진행하고 있다. 이들의 연구에 따르면 처음 300개의 단어의 사용이 특허문서의 다른 부분에 비해 가장 좋은 분류 결과를 나타냈으며, IPC 서브클래스 레벨에서 Top-prediction, Three-guesses, All-category의 세 가지 정확도 측정기준으로 분류성능을 평가했을 때, Top-prediction과 All-category 평가에서 SVM이 41%, 48%의 정확도로 가장 좋은 성능을 보이고, Three-guesses 평가에는 KNN이 62%로 가장 좋은 분류 결과를 보였다. 여기서 사용된 처음 300개 단어는 제목, 발명자, 특허권자, 요약, 명세서에서 추출한 처음 300단어로 구성되고 있다.

Larkey[5]의 연구에서는 US 특허를 대상으로 USPC 분류기준으로 특허문서의 분류를 수행하였다. 분류모델은 KNN을 사용하였으며, 제목, 요약, background, summary의 처음 20줄과 청구항 모두를 사용하여 분류 시에 성능이 가장 좋음을 보였다. 이들은 USPC의 395번 클래스의 2.09번 서브클래스 속하는 speech signal processing 서브클래스들에 대하여 성능 평가를 수행하였을 때, 25%~32%의 분류 정확도를 얻었다.

C.J. Fall 등과 Larkey는 특허문서의 특정 필드가 분류성능에 미치는 영향을 실험을 통해 보였으나 처음 300개의 단어, 처음 20줄의 사용에서 볼 수 있듯이 특허문서의 특징을 고려하여 특정 필드를 의식적으로 고려하고 있다고는 보기 어렵다.

* 본 논문의 특허명 뒤의 괄호안의 숫자는 특허 출원번호 임

(표 1) 특허문서 분류 관련 연구 정리
(Table 1) Summary of the related papers on patent classification

저자	IPC 분류 레벨	사용데이터	데이터 필드 구분	분류결과 (Best performance)
CJ Fall 외	클래스, 서브클래스	WIPO-alpha	제목 요약, 청구항, 처음 300단어	79%, 62%
Larkey	해당사항 없음 (USPC 395번클래스의 2.09 서브클래스에 속하는 speech signal processing 서브클래스)	USPTO	제목, 요약, background, summary의 처음 20줄, 청구항	32%
D. Tikk 외	클래스, 서브클래스, 메인그룹	WIPO-alpha	발명자, 특허권자, 제목, 요약, 청구항	85.56%, 75.05%, 55.58%
Y. Chen 외	서브그룹	WIPO-alpha	언급 없음	36.07%
Seneviratne 외	클래스, 서브클래스	WIPO-alpha	제목, 요약, 처음300단어	81%, 67%
박찬정 외	섹션	한국 특허문서	제목, 요약, 청구항	43%
김재호 외	해당사항 없음	일본 특허문서	기술분야, 목적, 해결방법, 청구항, 설명, 예	51.26%

또한 D. Tikk 외[6]는 계층적인 분류모델 HITEC을 제안하여 IPC 메인그룹 레벨까지의 분류를 시도 했으며 발명자, 특허권자, 제목, 요약, 청구항을 사용하였을 때 C.J. Fall의 세 가지 평가 기준에 따라 36~56%의 분류 결과를 보였다. 또한 C.J. Fall의 결과에 비해 클래스 레벨에서 약 6~10% 향상된 결과를 얻었으며 서브클래스 레벨에서 약 12~14% 향상된 결과를 보였다. Y. Chen 외[7]는 SVM과 KNN을 계층적으로 이용한 모델을 제시하여 IPC 서브그룹 레벨에서의 분류를 수행하였다. 이들은 기계학습 기법들을 결합한 모델을 통해 IPC 서브그룹 레벨에서의 분류를 시도하였으며 그 결과 약 36%의 분류 정확도를 얻었다. D. Tikk외 와 Y. Chen외 의 연구에서는 계층적인 기법을 활용한 모델을 통해 IPC 분류의 성능 향상을 보이고 있으나 특허문서의 구조적 필드에 대한 의미와 역할을 고려하지 않고 있다.

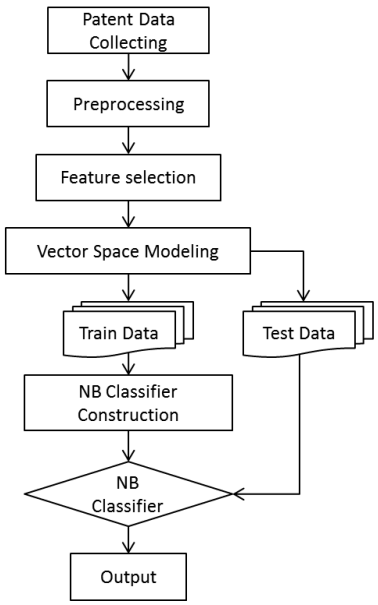
Seneviratne 외[8]는 특허문서 분류의 검색 시간과 scalability의 측면에서의 효율을 위하여 텍스트의 vector space modeling을 위해 주로 사용되어온 bag of words modeling의 대안으로 document signatures를 사용하는 방법을 제안하였다. 이들은 앞서 기술한 C.J. Fall의 연구의 영향을 받아 그들이 사용한 필드인 특허문서의 제목과 요약, 처음 300개의 단어로부터 document signature를 생성하는 방법을 제시하였다.

이 밖에도 특허문서 분류의 성능을 개선시키기 위한 방안으로 판별력 있는 특징을 추출하기 위한 연구도 수행되었다. 박찬정 외[9]는 Dominant Information 이라는 새로운 특징추출 방법을 제안하여 기존의 특징추출 방법들과 IPC 분류 성능에 끼치는 영향을 비교하였다. 이들은 한국 특허 약 1만7천 건을 대상으로 제목, 요약, 청구항을 사용하여 KNN 모델을 기반으로 하는 IPC 섹션 레벨에서의 분류를 시도하였다. 이들의 연구는 k=2, 10%의 특징 선택에서 43%의 정확도를 보인다. 하지만 IPC의 최상위 레벨인 8개의 섹션 레벨에서 분류를 수행하고 있어 실용성을 향상시키기 위한 연구가 요구되고 있다.

한편, 김재호 외[10]는 특허문서의 구조적인 특징을 고려하여 유사한 특허문서를 검색하기 위해 특허문서의 인덱싱 과정에 특허문서의 세분화된 필드를 사용하는 방법을 제안하였다. 이들은 일본 특허문서를 대상으로 하고 있으며 특허문서의 기술분야, 목적, 해결방법, 청구항, 설명 그리고 예 필드를 필드별로 교차 비교하는 과정에서 각 필드의 유사성을 계산하고 있다. 이들의 연구에서 기술분야, 목적, 청구항 영역이 다른 영역보다 유사문서 검색에 더 적합한 것으로 나타났다. 직접적인 IPC 분류 부여에 대한 연구는 아니지만 검색 과정에서 특허문서의 각 필드를 의식적으로 고려한 특허문서 검색을 수행한 연구의 하나라고 볼 수 있다. (표 1)은 이상의 관련연구들에 대하여 비교 분석한 내용을 요약한 것이다.

3. 기술분야와 배경기술 필드를 중심으로 한 특허문서의 IPC 분류

이 장에서는 IPC의 구조와 특허문서의 구조에 대하여 살펴본 후 본 논문에서 제안하는 특허문서의 구조적인 필드의 의미를 고려하여 '배경기술'과 '기술분야' 필드를



(그림 4) IPC 다중 분류 프로세스
(Figure 4) IPC multi-label classification process

사용한 특허문서 분류방법에 대하여 설명한다.

본 논문에서 제안하는 특허문서의 IPC 서브클래스 레벨에서의 분류는 데이터수집, 전처리, 특징단어 선택, 벡터 스페이스 모델링, 분류 모델 구축, 구축된 분류기를 통한 특허문서 분류의 단계로 수행된다. 이 과정은 (그림 4)와 같다.

특허문서의 분류를 위한 첫 번째 단계인 특허문서 수집과정에서는 분류에 필요한 데이터를 충분히 수집하기 위하여 Kipris plus[11]에서 제공하는 Open API를 사용한다. Kipris plus는 특허청이 개방중인 모든 특허정보를 실시간으로 활용할 수 있는 Bulk Data, Open API 방식의 서비스를 제공하고 있다. 이를 통해 2011년 1월 1일부터 2015년 12월 31일 까지 국내에 등록된 특허 전문 566,747건을 수집하였다. 수집된 pdf 형식의 특허문서로부터 분류에 영향력이 있는 대표적인 필드인 '제목', '요약', '청구항', '배경기술'과 '기술분야' 필드를 파싱한 뒤 이 중에서 '기술분야' 및 '배경기술' 필드가 없는 문서를 제외하여 564,793건의 특허문서 집합을 구축하였다. 이 특허문서 집합은 총 630개의 IPC 서브클래스로 분류되어 있다.

두 번째 단계에서는 수집된 특허문서에 대하여 전처리 과정을 수행한다. 목표클래스인 630개의 IPC 서브클래스에 대한 분류에 있어서 전처리는 매우 중요한 작업이다. 전처리에 따른 데이터의 정제 결과가 분류성능에 큰 영향을 끼칠 수 있기 때문이다. 이 단계에서 특허문서로부터 그 문서를 대표하는 최소 의미단위라고 볼 수 있는 명사를 추출한다. 이때 사용한 툴은 한국어 형태소분석기 KLT2000[12]이다. 또한 특허문서집합에는 특허문서의 내용에는 영향을 미치지 않는, 특허문서에서만 자주 쓰이는 용어들이 존재한다. 그 예로 '발명', '청구', '출원', '특허', '명세', '도면'과 같은 단어를 들 수 있다. stopword라고 일컫는 이러한 단어들을 효과적으로 제거하기 위하여 특허문서집합에 대한 stopwordlist를 구축한 뒤 전체 특허문서집합내의 각 특허문서에서 제거한다. 본 연구에서 사용한 stopwordlist는 지적재산권용어사전, 특허·실용신안 심사기준 그리고 특허용어사전으로부터 추출한 색인어를 기반으로 구축하였으며, 1860개의 stopword로 이루어져있다.

그 다음으로 특징 선택 과정에서는 본 연구의 목표 클래스인 630개의 IPC 서브클래스에 대하여 보다 판별력있는 feature를 선별한다. 이 과정에서 TF-ICF [7]를 사용한다. TF-ICF는 정보검색 분야의 대표적인 가중치 부여방법 중 하나인 TF-IDF 방식을 다중 분류에 적합하게 변형한 것이다. 이를 통해 각 IPC 서브클래스 별 특징단어들의 가중치를 계산하며 그 식은 아래와 같다.

$$TF_{ij} = avg(freq_{ij})$$
$$ICF_i = \log_2(N/n_i)$$
$$w_{ij} = TF_{ij} \times ICF_i$$

여기서 TF_{ij}는 IPC 서브클래스 분류 c_j에서의 특징단어 f_i의 평균빈도이며, ICF_i 식에서의 N은 IPC 서브클래스 분류의 개수, n_i는 특허문서집합에서 f_i를 포함하는 IPC 서브클래스 분류 c의 개수이다. 이로부터 IPC 서브클래스 c_j에서의 f_i의 가중치인 w_{ij}는 TF_{ij}와 ICF_i의 곱으로 계산된다. 이와 같이 TF-ICF 가중치 값에 따라서 각 IPC 서브클래스 별 상위 k개의 특징단어를 선택하여 특징단어 집합을 구축한다. 본 연구에서는 TF-ICF에 따른 특징단어들의 랭킹을 각 서브클래스별로 분석한 결과, k = 100 일 때 각 서브클래스의 특징단어들이 그 서브클래스 분야를 대표하면서 다른 분야와 잘 구분되는 것으로 판단하였다. 이에 따라서 각 IPC 서브클래스 별 상위 100개의 특징단어를 뽑아 중복은 제거한 뒤 총 33,047개의 단어로 이루어진 특징단어 집합을 구축하였다. (표 2)는 특징단어 집합의 구축에 사용된 TF-ICF 가중치 상위 100개의 특징단어들 중에서 30위까지를 A01클래스에 속하는 서브클래스에 대하여 내림차순으로 정리한 것이다. (표

(표 2) TF-ICF를 이용한 IPC 서브클래스 별 특징단어 추출 결과 (A01 클래스)
(Table 2) Extracted features for each IPC subclasses using TF-ICF (Class A01)

	A01B	A01C	A01D	A01F	A01G	A01H	A01J	A01K	A01M	A01N	A01P
1	써레	종자	예초기	베일러	재배	유전자	젖소	인공어초	해충	방제	방제
2	트랙터	파종기	예취	베일	화분	형질전환	착유컵	낚시	농약	균주	제초제
3	써레판	파종	콤바인	탈곡	비닐하우스	식물체	유두컵	낚싯대	퇴치	제초제	살충제
4	두둑	모종	예초	예취	식물	프로모터	치즈	수족관	방제	농약	조성물
5	트랙터용	비료	예취부	베일포장기	양액	식물	생유	낚싯줄	모기	살충제	식물병
6	작업기	이앙기	수확	사료작물	버섯	애기장대	착유	낚시대	방제기	식물병	살충
7	로터베이터	씨앗	예초기용	볏짚	식재	단백질	착유기	해삼	살포	식물	식물
8	배토기	발아	수확기	콤바인	수목	종자	요구르트	봉돌	약액	일아미노	해충
9	쟁기부	육묘상자	잔디	곡물	작물	벼	티트컵	애완동물	조류	조성물	농약
10	무논	퇴비	제초기	탈곡기	식생매트	재조합	유두	어류	야생동물	헥사하이드로프탈라진	방제제
11	쟁기	식부	칼날	롤베일	녹화	발현	진공호스	양식	약제	해충	추출물
12	파종기	씨앗필름	작물	베진	식생	서열	우유	집어등	포획장치	폐닐	잡초
13	고랑	이식기	잡초	콩	배지	plant	착유유닛	벌통	유인	미생물	작물
14	파종	볍씨	집초	급동	재배방법	벡터	지방구	낚시찌	포충기	옥소	미생물
15	경운기	마늘	집초기	예취부	온실	형질	양액	사육	까치	바실러스	화합물
16	씨앗필름	종순	작업봉	농산물	잘피	육종	진공센서	어초	살충제	살충	기피제
17	경운	멀칭필름	트랙터	작물	수경재배	신품종	마구간	사육수	둥지	화학식	균주
18	농기계	살포	땅속작물	건초	생육	al	유청	낚시줄	농작물	방제제	유효성분
19	써레판이	직파기	탈곡부	원형베일러	가로수	at	착유용	사료	유해동물	피리미딘	살미생물제
20	모종	살포기	곡간	베일링	농작물	품종	착유관	축사	살충	메톡시페닐	병해
21	쟁기날	직파	탈곡	목초	콩나물	세포	유축구	낚시바늘	모기향	화합물	살균제
22	종순	트랙터	농작물	사일로	발아	식물세포	착유정보	물고기	끈끈이	작물	생물농약
23	묘순	이식	예취날	곤포	생장	형질전환체	젖꼭지	가두리	곤충	메틸	흰가루병
24	이랑	두둑	수확장치	탈곡부	토양	투인원	착유실	양식장	퇴치기	잡초	방제방법
25	쟁기장치	비료살포기	예취기	곡립	상토	과발현	착유펌프	가축	수목	세포	선충
26	흙	파종장치	예초작업	조사료	종자	작물	착유우정보	원줄	살포장치	추출물	적조
27	베일러	상토	제초작업	래핑암	옹벽	생합성	여과챔버	미끼	분무기	병해	제초
28	이앙기	작물	모우어	사료	식물공장	코딩	균질단계	해파리	파리채	생물농약	활성
29	디봇	육묘	과일	급치	뿌리	비생물적	걸름망	수조	롤트랩	제제	항균
30	배토	점파	마늘	제어체	보온덮개	수확량	원유	해조류	멧돼지	활성	유해생물

2)의 A01B 열의 단어들은 써레, 트랙터, 작업기, 로터베이터와 같은 농기구들이 주를 이루고, A01C 열은 종자, 파종, 모종, 비료 등과 같이 파종과 관계된 단어들이, 그리고 A01P 열은 방제, 제초제, 살충제, 농약 등과 같이 농약과 관계된 단어들이 주를 이루는 것을 볼 수 있다. 이 단어들을 (표 3)의 IPC 분류 정의와 비교해 보면 TF-ICF 가중치 방법으로 선별한 특징단어들이 각각의 IPC 서브클래스를 잘 대표하고 있음을 볼 수 있다.

다음 단계인 벡터 스페이스 모델링 단계에서는 위의 과정에서 구축한 특징단어 집합을 기반으로 특허문서를 특징 벡터의 형태로 변환한다. 그리고 단순 빈도로 이루어진 특징 벡터들을 TF-IDF를 사용하여 가중치가 부여된 벡터로 변환한다.

특허문서 분류모델 구축단계에서는 630개의 목표클래스에 대하여 특허문서들을 다중 레이블로 분류하기 위한 모델을 구축한다. 예를 들어 특허 '식품 저장상태 감지용 센서에 무선으로 에너지를 공급하는 냉장고 (1020070015499)'는 제목만 으로부터 알 수 있듯이 F25D (냉장고; 냉각실; 아이스박스; 다른 서브클래스에 속하지 않는 냉각 또는 동결장치)와 A23B (식육, 어류, 난류, 과일, 채소, 식용종자의 보존; 과일 또는 야채의 화학적 숙성; 보존, 숙성 또는 통조림 제품)의 IPC를 가지며, 특허 '멀티미디어 전송 시스템에서 미디어 전송 패킷 전송 방법 및 장치 (10201300433855)'는 H04N (화상통신)과 H04L (디지털 정보의 전송)의 IPC를 가진다. 이와 같이 다중 레이블을 가질 수 있는 특허문서들을 분류하기위하여 본 연구에서는 텍스트 분류에서 사용되는 가장 대표적인 기계학습 모델 중 하나인 Multinomial Naive Bayes 모델[13]을 사용하고 있다.

(표 3) A01클래스의 서브클래스 (2015.01 ver.)
(Table 3) Subclasses in Class A01 (2015.01 ver.)

Section	A	생활필수품, 농업
Class	A01	농업; 임업; 축산; 수렵; 포획; 어업
Subclass	A01B	농업 또는 임업에 있어서의 토작업; 농기구 또는 기구의 부품, 세부 또는 부속구 일반
	A01C	식부; 파종; 시비
	A01D	수확;예취
	A01F	탈곡; 짚, 건초 또는 그와 유사한 것의 곤포;짚, 건초 또는 그와 유사한 것을 다발로 만들거나 묶기 위한 정치장치 또는 수동구; 건초, 짚 또는 그와 유사하는 것의 절단; 농업 수확물 또는 원예 수확물의 저장
	A01G	원예; 채소, 화훼, 벼, 과수, 포도, 호프 또는 해초의 재배; 임업; 관수
	A01H	새로운 식물 또는 그것들을 얻기위한 육종 처리; 조직배양기술에 의한 식물의 증식
	A01J	낙농제품의 제조
	A01K	축산; 조류, 어류, 곤충의 사육; 어업; 달리 분류되지 않는 동물의 사육 또는 번식; 새로운 동물
	A01L	동물의 장제(Shoeing of animals)
	A01M	동물의 포획, 덫을 놓아 잡기 또는 몰기; 유해한 동물 또는 유해한 식물의 구제장치
	A01N	인간, 동물 또는 식물의 본체, 또는 그것들의 부분 보존; 살생물제(Biocides); 유해생물 기피제 또는 유인제; 식물생장조절제
	A01P	화합물 또는 조성물의 살생물, 유해 생물 기피, 유해 생물 유인 또는 식물 생장 조절 활성

마지막으로 특허문서 분류단계에서는 Multinomial Naive Bayes 분류기를 사용하여 특허문서의 IPC 서브클래스를 예측한다. 분류모델 구축과 분류단계에서 사용한 Multinomial Naive Bayes 모델은 다음의 식과 같다.

$$P(c|f_1,f_2,...,f_n) \propto P(c)\prod_{i=1}^{n} P(f_i|c)$$

$$\hat{c} = \arg\max_c P(c)\prod_{i=1}^{n} P(f_i|c)$$

$$= \arg\max_c P(c)\prod_{i=1}^{n} \frac{N_{ci}+\alpha}{N_c+\alpha n}$$

여기서 N_{ci}는 분류 c에서의 feature i의 발생횟수, N_c는 분류 c 에서의 모든 feature의 발생횟수이다. 그리고 본 연구에서는 $\alpha = 1$인 Laplace smoothing을 사용하였으며 상위 5개까지의 분류 c를 제시하도록 하고 있다.

또한 위의 Multinomial Naive Bayes 분류 모델 구축 및 분류 단계에서는 10-fold cross validation 방법으로 전체 특허문서집합을 Training set과 Test set을 구분한 뒤 Training set를 사용하여 분류기 모델을 구축하고, Test set에 대하여 IPC 서브클래스 레벨에서의 다중 레이블 분류를 수행한다.

4. 실험 및 결과분석

이 장에서는 본 논문에서 제안한 '기술분야'와 '배경기술' 필드를 중심으로 한 특허문서의 IPC 분류 성능을 평가하고 기존 연구에서 주로 사용해온 제목, 요약, 청구항 필드를 대상으로 한 IPC 분류 성능을 비교하여 그 결과에 대하여 분석하고 고찰한다.

4.1 정확도 측정 기준

본 연구에서는 사람의 의사결정에 보조수단으로 활용될 수 있는 실용성 있는 분류기를 목적으로 한다. 따라서 특허문서의 각 필드를 대상으로 수행한 분류의 성능을 평가하기 위하여 다음과 같은 두 가지 정확도 척도를 사용한다. 싱글 매치 정확도는 분류기의 예측결과 중 어느 하나가 실제 문서의 분류와 일치하는 경우에 분류 성공이라고 간주한다. 그리고 올 매치 정확도는 분류기의 예측 결과가 실제 문서의 분류를 모두 포함하는 경우에 분류 성공이라고 간주한다. 이와 같은 방법은 분류기가 적절한 IPC를 추천해주는지에 대한 여부를 평가하기 위한 방법이며, C.J. Fall의 연구에서 사용된 메인 IPC 비교를 기반으로 하는 평가방법을 다중 레이블 추천을 위해 변형한 형태이다.

$$precision_{single} = \frac{1}{n}\sum_{i=1}^{n} I(T_i \cap P_i)$$

$$precision_{all} = \frac{1}{n}\sum_{i=1}^{n} I(T_i \subseteq P_i)$$

여기서 I는 indicator function을 나타내며, T는 특허문서의 실제 IPC 서브클래스 분류, P는 분류기가 예측한 IPC 서브클래스 분류이며 n은 Test set의 특허문서 개수이다.

4.2 각 필드별 분류 성능 분석

‘기술분야’와 ‘배경기술’ 필드가 분류성능에 미치는 영향을 살펴보기 위하여, 기존의 연구에서 주로 사용해온 제목, 요약, 청구항 필드를 가지고 다음과 같은 비교실험을 구성하였다. 특허문서의 (1) 제목, (2) 요약, (3) 청구항, (4) 기술분야 및 배경기술 필드만 사용한 경우에 대하여 각각 IPC 분류 성능을 비교하였다. 그리고 (5) 제목과 요약, (6) 제목, 요약, 청구항 필드를 함께 사용한 경우와 이들 (5), (6)에 대하여 ‘기술분야’와 ‘배경기술’ 필드를 추가하여 (7) 제목, 요약, 기술분야 및 배경기술 필드를 함께 사용한 경우, (8) 제목, 요약, 청구항, 기술분야 및 배경기술의 필드를 모두 사용한 경우로 나누어 IPC 분류를 수행하였다.

분류 실험은 Intel Xeon 8 core cpu, 128GB RAM, 64bit Linux OS상에서 수행되었으며, Multinomial Naive Bayes 분류기는 scikit-learn라이브러리를 이용하여 Python 환경에서 구현하였다.

분류에 앞서 8가지의 서로 다른 필드로 구성된 특허문서 집합에 대하여 각 문서의 평균 길이를 살펴보았을 때, (1) 제목 필드는 4단어, (2) 요약 필드는 48.19단어, (4) 배경기술 및 기술분야 필드는 141.60단어, (3) 청구항은 212.72단어이며 (8) 제목, 요약, 청구항, 기술분야 및 배경기술 필드를 모두 사용한 경우는 333.52단어로 나타났다. (표 4)는 이 결과를 나타내고 있다.

(표 4) 사용된 필드별 특허문서의 평균 길이
(Table 4) The average length of patent documents by fields

구분	단어 수
(1) 제목	4.00
(2) 요약	48.19
(3) 청구항	212.72
(4) 기술분야, 배경기술	141.60
(5) 제목, 요약	51.62
(6) 제목, 요약, 청구항	243.47
(7) 제목, 요약, 기술분야, 배경기술	182.60
(8) 제목, 요약, 청구항, 기술분야, 배경기술	333.52

이와 같은 8가지의 서로 다른 필드로 구성된 특허문서 집합에 대하여 IPC 분류를 수행하였다. 그 결과 (7) 제목, 요약, 기술분야 및 배경기술 필드를 사용한 경우 564,793개의 테스트 문서 중에서 492,513개의 문서에 대하여 분류성공을 보이며 87.20%의 정확도로 가장 높았으며, (4) 기술분야 및 배경기술이 86.60%, (8) 제목, 요약, 청구항, 기술분야 및 배경기술의 필드를 모두 사용한 경우가 85.67%로 나타났다.

반면에 (3) 청구항 필드만 사용한 경우 76.66%의 정확도를 보이며 분류성능이 가장 낮았으며 (2) 요약 필드만 사용한 경우가 77.57%, (1) 제목 필드만 사용한 경우가 78.24%의 정확도를 보였다.

또한 분류 성공 기준을 분류기가 문서의 실제 IPC 분류를 모두 제시한 경우로 한정하여 정확도를 측정하였을 경우에도 (7) 제목, 요약, 기술분야, 배경기술 필드를 사용한 경우가 70%로 가장 높은 정확도를 보이며, 기술분야 및 배경기술이 포함된 (4) 기술분야 및 배경기술과 (8) 제목, 요약, 청구항, 기술분야 및 배경기술의 경우가 각각 69%, 68.31%로 그 뒤를 이었다.

여기서 몇 가지 특허문서를 대상으로 각 필드의 사용에 따라 특허문서가 어떻게 분류되었는지 분류결과를 살펴보면 다음과 같다. 특허 ‘전지용 세퍼레이터, 및 전지용 세퍼레이터의 제조 방법 (1020147007370)’은 H01M(화학에너지를 전기에너지로 직접 변환하기 위한 방법 또는 수단, 예. 배터리), B32B(적층체, 즉 평평하거나 평평하지 않은 형상의 층으로 조립된 제품), C08J(마무리; 일반적 혼합 방법; 서브클래스 C08B, C08C, C08F, C08G 또는 C08H에 포함 되지 않는 후 처리), C08K(무기 또는 비고분자 유기 물질의 배합 성분으로서의 사용), C08L(고분자 화합물의 조성물)의 5개의 IPC 서브클래스를 가지는 문서이다. 이 문서의 (4)기술분야 및 배경기술, (7)제목, 요약, 기술분야 및 배경기술의 필드를 사용하여 분류하였을 때, 5개의 IPC 를 모두 잘 예측하였으며 (2)요약, (5)제목과 요약, (8)제목, 요약, 청구항, 기술분야 및 배경기술의 경우는 C08K를 제외한 4개, 그리고 (1)제목, (3)청구항, (6)제목, 요약, 청구항의 필드는 C08K, C08L을 제외한 3개의 IPC 서브클래스를 바르게 예측하였다.

그리고 특허 ‘동식물의 동식물성 기름을 이용한 침출수 처리장치 및 처리방법 (1020150122814)’은 7개의 IPC 서브클래스를 갖고 있으며 각각 C02F(물, 폐수, 하수 또는 오니(슬러지)의 처리), A47J(주방 장비, 커피 분쇄기, 향신료 분쇄기, 음료를 만드는 장치), A61L(재료 또는 물건을 살균하기 위한 방법 또는 장치 일반; 공기의 소독, 살균 또는 탈취), B01D(분리), B04B(원심분리기), B09B(고체 폐기물의 처리), F26B(고체원료 또는 고형물에서 액체를 제거하는 것에 의한 건조)와 같다. 이 문서의 경우, (7)제목, 요약, 기술분야 및 배경기술의 경우 5개의 IPC 서브클래스를 바르게 예측하였으며 (2)요약, (3)청구

(표 5) IPC subclass 레벨에서의 분류 정확도
(Table 5) Classification precision at IPC subclass level

구분	precision (single)	precision (all)
(1) 제목	78.24%	60.78%
(2) 요약	77.57%	59.40%
(3) 청구항	76.66%	58.65%
(4) 기술분야, 배경기술	86.60%	69.00%
(5) 제목, 요약	79.59%	61.62%
(6) 제목, 요약, 청구항	79.13%	61.18%
(7) 제목, 요약, 기술분야, 배경기술	87.20%	70.00%
(8) 제목, 요약, 청구항, 기술분야, 배경기술	85.67%	68.31%

항, (4)기술분야 및 배경기술, (5)제목과 요약, (6)제목, 요약, 청구항, (8)제목, 요약, 청구항, 기술분야 및 배경기술의 경우 4개, 그리고 (1)제목의 경우 3개를 바르게 예측하였다.

이상의 결과를 종합하면 IPC 서브클래스 레벨까지의 분류에서는 기술분야 및 배경기술 필드가 분류 정확도에 중요한 역할을 하고 있음을 알 수 있다. 그러나 특허문서 분류에 청구항 필드를 사용하는 것은 정확도를 떨어뜨리는 경향을 보이고 있으며, 나아가 청구항 필드의 문서당 평균 길이가 요약 필드에 비해 약 4배, 기술분야 및 배경기술 필드의 전체 길이의 약 2배 긴 점을 고려했을 때, 분류과정의 처리시간 측면에서도 비효율적인 것으로 보여진다. 하지만 청구항을 독립청구항과 종속청구항으로 세분하여 독립청구항만을 고려하면 다른 결과가 도출될 수 있다는 의견도 있어, 이런 내용을 포함하여 향후 연구를 진행하는 것은 의미가 있다고 사료된다.

또한 특허 '센서 (1020047020796)'의 경우와 같이 4개의 IPC 서브클래스를 가지고 있으나 8가지의 분류실험에서 모두 서브클래스의 5개의 예측이 어긋나는 경우도 있었다. 이 문서의 실제 IPC 서브클래스는 G01L(힘, 토오크, 일, 기계적 동력, 기계적 효율 또는 유체압력의 측정), G01H(기계적 진동 또는 초음파, 음파 또는 아음파의 측정), B82Y(나노 구조의 특별한 사용이나 적용; 나노 구조의 측정이나 분석; 나노 구조의 제조나 처리), G01B(길이, 두께 또는 유사한 직선치의 측정; 각도의 측정; 면적의 측정; 표면 또는 윤곽의 불규칙성 측정) 로 분류되어 있으나 분류실험의 결과, G01R(전기변량의 측정; 자기변량의 측정), G06F(전기에 의한 디지털 데이터 처리) 또는 H04L (디지털 정보의 전송), H04W(무선통신네트워크)와 같은 서브클래스가 예측되었다. 실제 특허문서로부터 그 이유를 살펴보았을 때, 이는 특허의 제목인 '센서'와 같이, 적용되는 분야가 다양한 단어로 구성된 필드의 사용은 세분화된 서브클래스를 판별하기에 부족한 점과, 기술분야 및 배경기술, 요약, 청구항 필드 뿐만 아니라 본 연구에서는 사용하지 않은 발명의 상세한 설명 필드까지 포함하는 특허문서의 대부분의 내용을 기반으로 세부분야가 구분되는 경우도 존재할 수 있기 때문으로 사료된다. 때문에 세분화된 분야를 구별하는 성능의 향상과 향후 IPC의 최하위 레벨까지의 분류를 위해서는 특허문서의 전체 구조를 바탕으로 데이터 필드 선택에 대한 심화된 연구가 필요하다고 생각된다.

또한 본 연구에서는 Multinomial Naive Bayes 분류기를 기반으로 특허문서를 분류하였으나 SVM, KNN과 같은 다양한 모델을 적용하면 분류성능을 보다 향상시킬 수 있을 것으로 기대된다.

5. 결 론

본 논문은 특허문서에 대하여 효과적인 IPC 분류를 위하여 기계학습 기법 등과 같은 분류 모델에 중심을 두는 것이 아니라 특허문서의 데이터 구조 그 자체에 보다 집중하는 분류 방법을 제안하였다. 특허문서의 데이터 특징과 IPC 분류구조의 분석을 통해 IPC 서브클래스 레벨에서의 분류 수행에 있어서 의미있는 필드로 기술분야 및 배경기술 필드를 제시하였으며, 이를 확인하기 위해 특허문서의 구조적 필드가 IPC 서브클래스 레벨까지의 분류에 미치는 영향을 살펴보기 위한 특허문서의 각 필드의 조합으로 구성한 비교실험을 수행하였다. 그 결과 제목, 요약, 기술분야 및 배경기술 필드의 사용이 가장 높은 분류 정확도를 보였으며, 기술분야 및 배경기술 필드를 사용한 경우가 다른 필드의 사용에 비해 높은 분류 정확도를 보임으로써 기술분야 및 배경기술 필드가 IPC 분류에 있어서 중요한 역할을 하고 있음을 확인하였다. 또한 특허문서의 청구항 필드는 IPC 서브클래스 레벨까지의 분류에서 정확도를 떨어뜨리는 결과를 가져오는 사실도 확인하였다. 그리고 630개의 IPC 서브클래스 레벨에서의 다중 레이블 분류를 통하여 특허문서 분류의 현장으로의 적용 가능성을 확인하였다.

향후 연구로는 사람에 의한 특허문서의 분류과정에서 제목, 요약, 배경기술 및 기술분야 등의 포괄적인 내용을 활용하여 IPC분류의 상위 레벨의 분류를 확인하고, 청구항 등을 기반으로 최종적인 IPC분류를 결정하고 있는 점

을 고려해서, IPC 분류구조의 최하위 레벨인 서브그룹 까지의 실용적인 IPC 다중 레이블 분류를 목표로 특허문서의 각 필드의 특성을 고려하여 IPC분류 과정의 각 단계별로 특허문서의 필드를 선택적으로 선정하여 활용하는 계층적인 기계학습 모델 기반의 연구가 수행될 필요가 있다.

참 고 문 헌(Reference)

[1] "Intellectual Property Statistics for 2014," Korean Intellectual Property Office, ISSN 2092-5417, 2015.

[2] International Patent Classification Guide, http://www.kipo.go.kr/kpo/user.tdf?a=user.html.HtmlApp&c=40304&catmenu=m06_07_02_05&year=2015&ver=01

[3] "Guidelines for Examination," Korean Intellectual Property Office, ISSN 2092-8866.

[4] C.J. Fall, A. Torcsvari, K.Benzineb, G. Karetka, "Automated Categorization in the International Patent Classification," In ACM SIGIR forum, April 2003, vol. 37(1), pp. 10-25. http://dx.doi.org/10.1145/945546.945547

[5] LS. Larkey, "A Patent Search and Classification System," In the 4th ACM Conference on Digital Libraries, pages 19-87, Berkeley, CA, August 99. http://dx.doi.org/10.1145/313238.313304

[6] D. Tikk, G. Biró, A. Törcsvári, "A Hierarchical Online Classifier for Patent Categorization," In Emerging Technologies of Text mining: Techniques and Applications (2007), pp. 244-267. https://doi.org/10.4018/9781599043739.ch012

[7] Y.-L. Chen, Y.-C. Chang, "A three-phase method for patent classification," Information Processing and Management, Vol. 48, no. 6, pp. 1017-1030, 2012. https://doi.org/10.1016/j.ipm.2011.11.001

[8] D. Seneviratne, S. Geva, G. Zuccon, and G. Ferraro, "A Signature Approach to Patent Classification," Information Retrieval Technology Vol. 9460, pp. 413-419, 2016. https://doi.org/10.1007/978-3-319-28940-3_35

[9] C. Park, K. Kim, and D. Seong, "Automatic IPC Classification for Patent Documents of Convergence Technology Using KNN," Journal of KIIT. Vol. 12, no. 3, pp. 175-185, Mar. 2014. https://doi.org/10.14801/kiitr.2014.12.3.175

[10] J. Kim, K. Choi, "Patent Document Categorization based on Semantic Structural Information," In Proc. of the 17th Annual Conference on Human and Cognitive Language Technology, pp. 28-34, 2005. http://www.dbpia.co.kr/Article/NODE01065130

[11] KIPRIS (Korea Intellectual Property Rights Information Service) plus, http://plus.kipris.or.kr/

[12] KLT2000, Korean Morphologigal Analyzer, http://nlp.kookmin.ac.kr/

[13] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," In Seventh Australian joint conference on artificial intelligence, Springer, Berlin, pp. 488 – 499, 2004. https://doi.org/10.1007/978-3-540-30549-1_43

◐ 저 자 소 개 ◑

임 소 라(Sora Lim)

2010년 한국항공대학교 정보통신공학과(공학사)
2012년 한국항공대학교 대학원 정보통신공학과(공학석사)
2012년~현재 한국항공대학교 대학원 정보통신공학과 박사과정
2011년~현재 한국항공대학교 차세대방송미디어기술연구센터 연구원
관심분야 : 정보검색, 데이터 마이닝, 빅데이터 분석
E-mail : ebbunsora@kau.ac.kr

권 용 진(YongJin Kwon)

1986년 한국항공대학교 항공전자공학과(공학사)
1990년 일본 교토대학 대학원 정보공학과(공학석사)
1994년 일본 교토대학 대학원 정보공학과(공학박사)
1994년~현재 한국항공대학교 항공전자정보공학부 정교수
2007년~현재 경기도지역협력연구센터(GRRC) 센터장
관심분야 : 정보검색, mixed reality, 빅데이터 활용
E-mail : yjkwon@kau.ac.kr