ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2018 September: 24(3): 221~241 http://dx.doi.org/10.13088/jiis.2018.24.3.221

한국표준산업분류를 기준으로 한 문서의 자동 분류 모델에 관한 연구*

이재성
과학기술연합대학원대학교
과학기술경영정책학과
(*jslee@kisti.re.kr, jaeseong.lee@ust.ac.kr*)

전승표
한국과학기술정보연구원 데이터분석본부/
과학기술연합대학원대학교 과학기술경영정책학과
(*spjun@kisti.re.kr*)

유형선
한국과학기술정보연구원 데이터분석본부/
과학기술연합대학원대학교 과학기술경영정책학과
(*hsyoo@kisti.re.kr, hsyoo@ust.ac.kr*)

지식사회에 들어서며 새로운 형태의 자본으로서 정보의 중요성이 강조되고 있다. 그리고 기하급수적으로 생산되는 디지털 정보의 효율적 관리를 위해 정보 분류의 중요성도 증가하고 있다. 본 연구에서는 기업의 기술사업화 의사결정에 도움이 될 수 있는 맞춤형 정보를 자동으로 분류하여 제공하기 위하여, 기업의 사업 성격을 나타내는 한국표준산업분류(이하 'KSIC')를 기준으로 정보를 분류하는 방법을 제안하였다. 정보 혹은 문서의 분류 방법은 대체로 기계학습을 기반으로 연구되어 왔으나 KSIC를 기준으로 분류된 충분한 학습데이터가 없어, 본 연구에서는 문서간 유사도를 계산하는 방식을 적용하였다. 구체적으로 KSIC 각 코드별 설명문을 수집하고 벡터 공간 모델을 이용하여 분류 대상 문서와의 유사도를 계산하여 가장 적합한 KSIC 코드를 제시하는 방법과 모델을 제시하였다. 그리고 IPC 데이터를 수집한 후 KSIC를 기준으로 분류하고, 이를 특허청에서 제공하는 KSIC-IPC 연계표와 비교함으로써 본 방법론을 검증하였다. 검증 결과 TF-IDF 계산식의 일종인 LT 방식을 적용하였을 때 가장 높은 일치도를 보였는데, IPC 설명문에 대해 1순위 매칭 KSIC의 일치도는 53%, 5순위까지의 누적 일치도는 76%를 보였다. 이를 통해 보다 정량적이고 객관적으로 중소기업이 필요로 할 기술, 산업, 시장 정보에 대한 KSIC 분류 작업이 가능하다는 점을 확인할 수 있었다. 또한 이종 분류체계 간 연계표를 작성함에 있어서도 본 연구에서 제공하는 방법과 결과물이 전문가의 정성적 판단에 도움이 될 기초 자료로 활용될 수 있을 것으로 판단된다.

주제어 : 문서자동분류, 한국표준산업분류, 텍스트마이닝, 벡터공간모델, 자연어 처리

논문접수일 : 2018년 5월 31일 논문수정일 : 2018년 7월 9일 게재확정일 : 2018년 8월 20일
원고유형 : 일반논문(급행) 교신저자 : 유형선

1. 서론

지식사회라고 일컬어지는 21세기에 들어, 지식의 근간이 되는 정보는 새로운 형태의 자본으로서 그리고 일종의 에너지로서 그 중요성이 강조되고 있다(Drucker, 1993). 정보는 모든 경제활동의 효율성을 향상시키는 소중한 자원으로 활용될 수 있어, 각 기술사업화 주체는 이를 수집하고 생산하며 공유하지만 한편으로는 보호하고 통제하기도 한다. 정보는 돈과 같은 다른 자원에 비해 쉽게 생성되고 유통될 수 있다. 여기에는 기술의 발전이 큰 역할을 하고 있는데, 컴퓨터, 인터넷의 보급과 최근 급격히 성장하고 있는 소셜 네트워크 서비스로 인해 인간이 생산한 디지

* 이 연구는 한국과학기술정보연구원 주요사업의 지원으로 이루어졌습니다.

털 정보의 양은 기하급수적으로 늘어나고 또한 확산되고 있다. 그로 인해 최근에는 오히려 정보 과잉 현상을 겪게 되면서, 자신에게 꼭 필요한 정보를 확보하는데 많은 시간과 비용이 소요되고 있다(Gudivada et al., 1997). 이러한 비효율성은 정보의 분류 문제와 관련된다. 사용자에게 필요한 정보인지, 관심 있는 주제와 관련된 정보인지, 특정 분류체계 하에서 어느 분야에 해당하는 정보인지를 분류하는 것은 정보의 효율적 관리에 있어 매우 중요하기 때문이다.

정보의 분류는 기술사업화의 핵심 주체인 기업의 의사결정에도 중요한 요소로 작용한다. 사업 환경이 기업의 성과에 미치는 속도가 빨라지고 외부 환경 역시 급변하고 있기 때문에 사업을 하는 기업의 입장에서는 특별히 시의적절하고 신뢰성 있는 맞춤형 정보가 더욱 필요한 실정이다. 따라서 정보의 정확한 분류가 기업에게 필요한 정보를 신속히 확보하는데 도움이 된다. 특히 우리나라 기업체의 99%을 차지하고 있는 중소기업(Ministry of SMEs and Startups, 2014)의 경우에는 고품질의 정보를 획득하기 위한 내부 역량이 상대적으로 부족해 심각한 정보불균형을 초래하고 있다(National Information Society Agency, 2016). 따라서 이러한 중소기업의 기술사업화 의사결정에 도움이 될 수 있는 맞춤형 정보를 자동으로 분류하여 제공한다면, 정보의 과잉과 불균형으로 인한 문제를 어느 정도 해소해 줄 수 있을 것으로 기대된다.

한편 모든 기업체가 영위하는 사업은 그 성격에 따라 한국표준산업분류(이하 KSIC; Korea Standard Industry Classification)를 기준으로 분류할 수 있다. KSIC는 국내 산업 관련 통계 자료의 정확성, 비교성을 확보하기 위해 통계청에서 작성한 산업 분류 체계이다(Statistics Korea, 2008). KSIC는 생산 주체가 수행하고 있는 생산활동을 유사성에 따라 유형화 한 것으로 2007년 9차 개정안에 따르면 21개 대분류와 1,145개 세세분류로 구분된다. 따라서 특정 기업에 대한 맞춤형 정보 혹은 기업이 관심을 가질만한 정보는 그 기업이 속한 KSIC를 기준으로 분류해 볼 수 있다. 이러한 맥락에서 방대한 디지털 정보를 KSIC를 기준으로 분류하여 제공할 수 있는 방법이 제공된다면 기업의 기술사업화 의사결정에 큰 도움이 될 수 있다.

정보 혹은 문서 자동분류에 대한 연구는 대체로 기계학습에 기반을 두어 많이 이루어져 왔다(Yang, 1999; Sebastiani, 2002). 하지만 문서를 KSIC를 기준으로 분류하는 연구는 거의 진행된 바가 없는데, 이는 기계학습 기법을 기반으로 하는 모델을 구축하기 위해서는 미리 분류된 많은 양의 학습 데이터가 필요하지만 KSIC를 기준으로 분류된 학습 데이터가 거의 없기 때문이다. 이에 본 연구에서는 문서 간 유사도를 계산하는 방식으로 KSIC를 기준으로 문서를 분류하는 방법과 모델을 제시하였다. 구체적으로 KSIC 각 코드별 설명문을 수집하고 벡터 공간 모델을 이용하여 분류 대상 문서와의 유사도를 계산하여 가장 적합한 KSIC 코드를 제시하는 방법과 이를 구현하는 모델을 제시하였다. 그리고 국제특허분류(이하 IPC; International Patent Classification)의 데이터를 수집한 후 KSIC를 기준으로 분류하고, 이를 특허청에서 제공하는 KSIC-IPC 연계표와 비교함으로써 본 방법론을 검증하였다.

이어지는 제 2장에서는 문서 자동 분류와 가중치 부여 이론에 관한 선행연구를, 제 3장에서는 실험 모델과 사용 데이터 및 방법론에 대한 내용을, 제 4장에서는 실험을 통한 결과와 토의를, 마지막 제 5장에서는 시사점 및 기여와 함께

한계점과 향후 연구 방안에 대해 기술하였다.

2. 선행연구

2.1 자동 문서 분류 모델에 대한 연구

정보검색 모델은 질의 요구에 적합한 문서들을 사용자에게 제공함으로써 대용량의 데이터로부터 주어진 시간 내에 원하는 정보를 발견할 수 있도록 도와준다(Salton, 1987). 이전에는 이것을 모두 수작업으로 분류하였기 때문에 많은 비용이 수반될 수밖에 없었다(Lee, 1994). 하지만 문서 자동 분류는 대량의 문서를 효율적으로 관리하고 검색할 수 있게 하는 동시에 방대한 양의 수작업을 감소시키는 기능을 수행한다(Lee, 1994). 문서의 자동 분류는 1960년 대부터 정보검색의 한 분야로 연구되어지기 시작하여, 1980년대 말에 이르러 주요 이론적인 기초연구가 활발히 진행되었다. 이를 응용한 모델 관련 연구도 전문가가 수작업을 통해 생성해 낸 규칙을 기반으로 문서를 자동 분류하는 방법으로 주로 구현이 되어왔다(Jeong, 1993; Sebastiani, 2002).

전문가의 정성적 평가를 통한 규칙기반 문서 자동 분류의 연구는 주로 여러 기계학습 기법을 기반으로 이루어 지고 있다(Yang and Liu, 1999; Yang 1999; Sebastiani, 2002; Hong et al., 2014). 기계학습 기법은 도메인 지식에 독립적이며 대량의 정보를 다룰 수 있는 자동 분류 모델 구축을 가능하게 한다. 따라서 문서 자동 분류 모델은 도메인 지식에 의존한 전문가 모델 또는 지식기반 모델이 방대한 정보들을 처리해 내기 힘든 기존의 한계로 인해서 다양한 분야의 연구자들로부터 많은 관심을 받고 있다.

기계학습 기법을 기반으로 하는 자동 분류 모델에 대한 기존의 선행 연구들은 다음과 같다. 의학분야의 관련 문서의 주제명을 자동부여하기 위한 목적으로 자동 분류 모델이 연구된 바 있으며(Ruiz and Scrinivasan, 1999), 생명정보학 분야 논문의 색인어를 부여하기 위해 연구되기도 했다(Chang, 2000). 그밖에 뉴스 기사나 SNS 등의 웹 페이지 문서를 자동으로 분류하기도 했으며(Lewis and Gale, 1994; Craven et al., 1998; Craven et al., 2000; Kim and Kim, 2013; Lee et al., 2016), 사용자의 기호를 자동 분류의 결과에 따라 모델에 학습시키고(Pazzani and Billsus, 1996; Lang, 1995; Lee et al., 2007; Jeon and Choi, 2010), 전자메일을 자동으로 분류하거나(Lewis and Knowles, 1997; Kim, 2016), 개인 홈페이지를 찾을 때와(Shavlik and Eliassi-Rad, 1998; Park et al., 2000), 책을 추천하는 데에 응용되기도 하였다(Mooney and Roy, 2000; Choi, 2016).

2.2 벡터 공간 모델에 대한 연구

일반적으로 벡터 공간 모델은 불리언(boolean) 모델, 벡터(vector) 모델과 확률(probability) 모델로 구분되는 정보검색모델 중 하나의 기법을 일컫는다(Salton, 1989; Ponte and Bruce, 1998; Vapnik, 1998; Yang, 1999). 위의 모델들은 문서 표현방법과 가중치 할당 방법, 그리고 문서 Di와 질의 Q 의 유사도 계산방식에서 차이가 있다. 이 중에서도 벡터 공간 모델은 정보검색 기법 중 가장 널리 사용되고 있는 모델이다(Cooper, 1983; Radecki, 1988). 벡터 공간 모델은 단순하고 빠르기 때문에 오늘날 내용기반으로 설계된 웹 정보 검색에서 가장 전형적인 모델로 간주되고 있다(Dillon, 1983). 그리고 벡터 공간 모델은

각 문서를 그 문서가 포함하고 있는 색인 단어의 벡터로 나타낼 수 있으며, 문서 간 유사도는 벡터에 위치한 단어들 간의 거리로 계산할 수 있다는 특징을 가지고 있다(Salton et al., 1975).

벡터 공간 모델을 이용한 선행 연구로는 다음과 같다. Chang(2013)은 그래프에 기반한 텍스트 표현 모델을 제안하며 벡터 공간 모델과 비교하는 연구를 진행한 바 있다. 그리고 Kim and Chang(2014)은 텐서 공간 모델을 개량한 3차 텐서(3-order tensor) 모델을 제안하고 그 성능을 벡터 공간 모델과 비교하여 검정하였다. Hamedani and Kim(2014)은 벡터 공간 모델과 확률 모델을 비교하며 실제 학술 논문의 유사도 계산에 있어 벡터 공간 모델이 더 우수했음을 증명했다. 그리고 Lee(2017)은 한국 농촌 계획 협회(KSRP) 간행물의 연구 동향을 정량적으로 분석하기 위해 UN의 '지속 가능한 개발 목표(SDGs)'의 17개 SDG와 771개의 간행물 데이터로 벡터 공간 모델을 사용한 바 있다. 그 밖에도 고객의 목소리, 사용자의 리뷰 등을 분석하는데 활용되어 왔다(Kim and Jung, 2013; Byun et al., 2016).

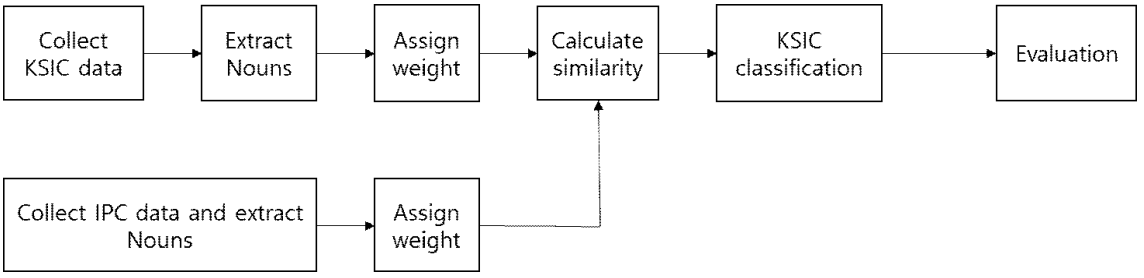
일련의 선행연구와 같이 벡터 공간 모델을 사용하기 위해서는 문서의 벡터 공간에 있는 단어의 가중치를 계산해야 한다. 이때, 단어의 가중치 기법을 달리 적용함으로써 벡터 공간 모델의 검색 성능을 향상시킬 수 있다. 왜냐하면 가중치 계산 방법은 문서들의 정보량에 따라 상대적인 중요도를 나타내기 때문에 전체적인 모델의 분류 정확도에 영향을 끼치기 때문이다. 이러한 가중치 계산에는 여러가지 방식이 존재하는데, 그 중 TF(Term Frequency)와 TF-IDF(Term Frequency-Inverse Document Frequency)가 대표적이다. TF는 가장 직관적인 방법으로, 단순하게 문서 내에 출현하는 해당 단어의 총 빈도수를 유사도 계산에 필요한 가중치로 사용한다. 하지만 조사와 같이 자질의 특성이 매우 낮은 단어가 과도하게 높은 값을 할당 받아 벡터 공간 모델의 성능저하를 유발한다. 한편 TF-IDF는 자질 특성을 제대로 반영하지 못하는 TF의 단점을 보완한 방법이다. 실제로 TF-IDF는 텍스트를 기반으로 하는 정보 분류 연구의 약 83%에서 쓰이고 있다(Beel et al., 2016). 이를 통해 질의 조건에 근접한 문서 색인이 가능해지며, 거리 측정 기법 즉 Cosine 순위화 등의 기법이 문서들을 질의어를 기준으로 정렬해 줄 수 있다는 특징을 갖는다.

특히 질의어를 사용자 정보 요구 형태에 따라 상대적인 중요도를 매겨 분류 결과에 대한 순위를 쉽게 부여 할 수 있다는 점이 벡터 공간 모델이 갖는 강력한 특징이다. 왜냐하면 사용자의 질의와 가장 유사도가 높은 순위의 문서부터 우선적으로 검토할 수 있기 때문에, 필요한 정보를 얻는데 소모되는 시간을 최소화 할 수 있기 때문이다(Lee, 1994). 이러한 특징은 페이지랭크 알고리즘을 적용한 세계 최대의 검색엔진인 구글에 의해 정보검색에 있어서 얼마나 많은 문서를 데이터 베이스에 가지고 있는지 보다, 가지고 있는 문서를 얼마나 잘 순위를 매기어 주느냐가 가장 중요하다는 사실로 이미 증명된 바 있다(Witten et al., 1999).

3. 연구모형

3.1 실험 데이터 및 연구 방법론

<Figure 1>은 본 연구의 전체적인 수행 프로세스를 나타낸다. 문서를 KSIC로 분류하기 위해 먼저 KSIC 설명문 데이터를 확보하여 TF-IDF



〈Figure 1〉 Research Model Design

계산을 통해 가중치를 부여하였다. 분류대상문서로 선택한 IPC 설명문 데이터에 대해서도 같은 과정을 거치고, 서로 간의 유사도 를 Cosine 내적 각의 크기를 계산함으로써 적합한 KSIC로 분류한 후, 이를 검증하였다. 본 연구를 진행하는데 사용된 실험 환경은 다음과 같다. Ubuntu Linux 16.04 OS의 R ver3.4.3로 모델을 구축하였고, 사용된 R 패키지로는 XML, rvest, KoNLP, tm, NIADic, stringr, slam 등이 있다.

본 연구에서 제안하는 자동 분류 모델에서 분류 기준 데이터베이스로 사용될 문서(Document) 집합으로 2007년에 9차 개정된 KSIC 설명문 데이터를 수집하였다. KSIC는 대분류부터 세세분류까지 5단계로 구분되는데, 9차 개정된 KSIC 기준 5자리의 세세분류 코드는 총 1,145개에 달한다. 통계청은 각각의 코드별로 해당 분류를 설명하는 설명문을 제공하는데, 이를 확보하기 위해 통계청이 운영하는 통계분류포털의 KSIC 검색 페이지를 HTML 소스코드 파싱(Parsing)으로 가져오도록 웹 크롤러(Web Crawler)를 디자인했다. 이러한 과정을 통해 얻어진 예시(KSIC A01121 코드) 결과는 아래 <Table 1>과 같다.

자동 분류 모델을 통해 KSIC를 기준으로 분류될 문서는 각종 보고서와 뉴스 기사, 그리고 SNS 등 다양한 유형이 있을 수 있다. 하지만 특별히 본 연구에서는 제안하고자 하는 모델의 분류 정확도 검증을 위해 분류의 대상이 되는 문서 즉 질의(Query)될 문서 집합으로 IPC 데이터를 사용

〈Table 1〉 KSIC's classification system and the example of code(A01121)

Label	Code	Index	Definition statement
Section	A	농업	노지에서 각종 채소작물을 재배하는 산업활동을 말한다. 노지에서 깻잎 등과 같이 채소로 사용하기 위하여 각종 작물을 재배하는 경우에도 여기에 분류된다. 풋마늘 및 풋고추 재배 파 및 양파 재배 아스파라거스 재배 참외 수박 및 멜론 재배 잎 및 열매 채소 재배 가지 재배 강낭콩 재배 고추냉이재배 고추노지재배 고추재배 깻잎 재배 나물 재배 농업 당근 재배 도라지 재배 마늘 재배 멜론 재배 무 재배 미나리 재배 배추 재배 부추 재배 뿌리채소 재배 상추 재배 수박 재배 순무 재배 시금치 재배 아스파라거스 재배 양배추 재배 양엉겅퀴 재배 양파 재배 열매 채소 재배 오이 재배 잎 및 열매 채소 재배 잎줄기 채소 재배 잎채소 재배 줄기 채소 재배 쪽파 재배 참외 재배 채두류 재배 채소작물재배 채소 작물 재배 초본성 과실 재배 취나물 재배 치커리 재배 토마토 재배 파 재배 풋고추 재배 풋마늘 재배 호박 재배
Division	01	농업	
Group	1	작물 재배업	
Class	2	채소, 화훼작물 및 종묘 재배업	
Sub-Class	1	채소 작물 재배업	

하였다. IPC는 다양한 기술분야에 걸쳐 수 많은 특허들이 출원되고 등록되고 있는 것을, 조사 및 검색하거나, 또는 심사할 때에 특허의 기술 내용에 따라서 이를 분류를 위해 WIPO(World Intellectual Property Organization)에 의해 작성되었다(WIPO, 2017). IPC 분류체계의 초기 목적은 다량의 특허를 보다 체계적으로 분류해 효과적으로 검색하기 위해서였지만, 현재에는 선행기술조사, 기술동향분석, 특허심사 등에 유용한 자료로써 활용되고 있다. IPC 분류체계는 섹션부터 서브그룹까지 5단계로 구분되는데, 본 연구에서 사용한 서브클래스 단위에서는 총 643개의 코드로 구성된다.

IPC 데이터는 KSIC와 마찬가지의 방법으로 수집하였다. IPC는 특별히 설명문이 정의되어 있지는 않지만 IPC 코드의 명칭이 비교적 상세한 편이다. 따라서 본 연구에서는 서브클래스의 명칭과 해당 서브클래스에 속하는 하위 단위인 메인그룹, 서브그룹의 명칭을 모두 합하여 서브클래스 단위의 설명문으로 활용하였다. 특허청에서 제공한 IPC 조회 프로그램으로부터 분류코드별 명칭 데이터를 확보하였으며, 이러한 결과를 통해서 얻어진 예시 결과는 <Table 2>와 같다.

이렇게 확보된 각각의 KSIC와 IPC의 설명문 데이터는 유의미한 자질의 명사 단어 집합을 추출하기 위해서 많은 정제 작업을 거쳐야만 했다. 본 연구에서는 일련의 작업을 수행하기 위해 한국과학기술정보연구원에서 개발한 KnowledgeMatrix Plus(이하 KM+)를 활용했다. KM+는 텍스트 마이닝과 네트워크 분석에 수반되는 텍스트 데이터 정제, 형태소 분석, 행렬 데이터 생성 등 데이터 전처리를 지원하는 분석 툴이다. KM+의 주요기능으로는 한글 형태소 분석을 위한 루씬의 Arirang 형태소 분석기가 탑재되어 있다. 따라서 본 연구에서는 이를 이용하여 각각의 KSIC와 IPC의 설명문 데이터로부터 실험에 필요한 단어-문서 행렬(Term-Document matrix)을 생성하였다. 그리고 KM+를 이용하여 자체적으로 설정한 시소로스(thesaurus)를 적용하여 스테밍(stemming), 불용어(stop-words) 처리 등의 작업을 수행하였다.

이후 각각의 KSIC와 IPC의 설명문 데이터에서 추출된 명사 단어 집합의 출현 빈도와 문서 빈도를 계산하는 작업을 수행하였다. 이때 TF-IDF의 다양한 계산법을 통해 단어-문서 행렬(Term-Document Matrix)의 레코드에 가중치를 부여하였다. 가중치의 계산은 R의 tm라이브러리에서 제공하는 weightSMART 계산 모듈을 이용하였다. weightSMART는 Gerald Salton에 의해

〈Table 2〉 IPC's classification system and the example of code(A01B)

Label	Code	Index	Definition statement
Section	A	생활필수품 농업	농업 또는 임업에 있어서의 토작업 농기구 또는 기구의 부품 세부 또는 부속구 일반 수 작업구 고정날이 있는 쟁기 구동되지 않는 회전구가 있는 쟁기 디스크 쟁기 쟁기 써레 그와 유사한 형태로 사용할 수 있는 디스크 형상의 토 작업구 구동 회전구가 있는 쟁기 진동 굴취 구멍 뚫는 기구가 있는 쟁기 특별한 목적을 쟁기 유사한 농작업기 쟁기의 세부 기구 요소 특수한 부가장치가 있는 쟁기… (생략)
Class	01	농업; 임업; 축산; 수렵; 포획; 어업	
Sub-Class	B	농업 또는 임업에 있어서의 토작업; 농기구 또는 기구의 부품, 세부 또는 부속구 일반	
Main-Group	1	수(手)작업구	
Sub-Group	02	가래; 삽	

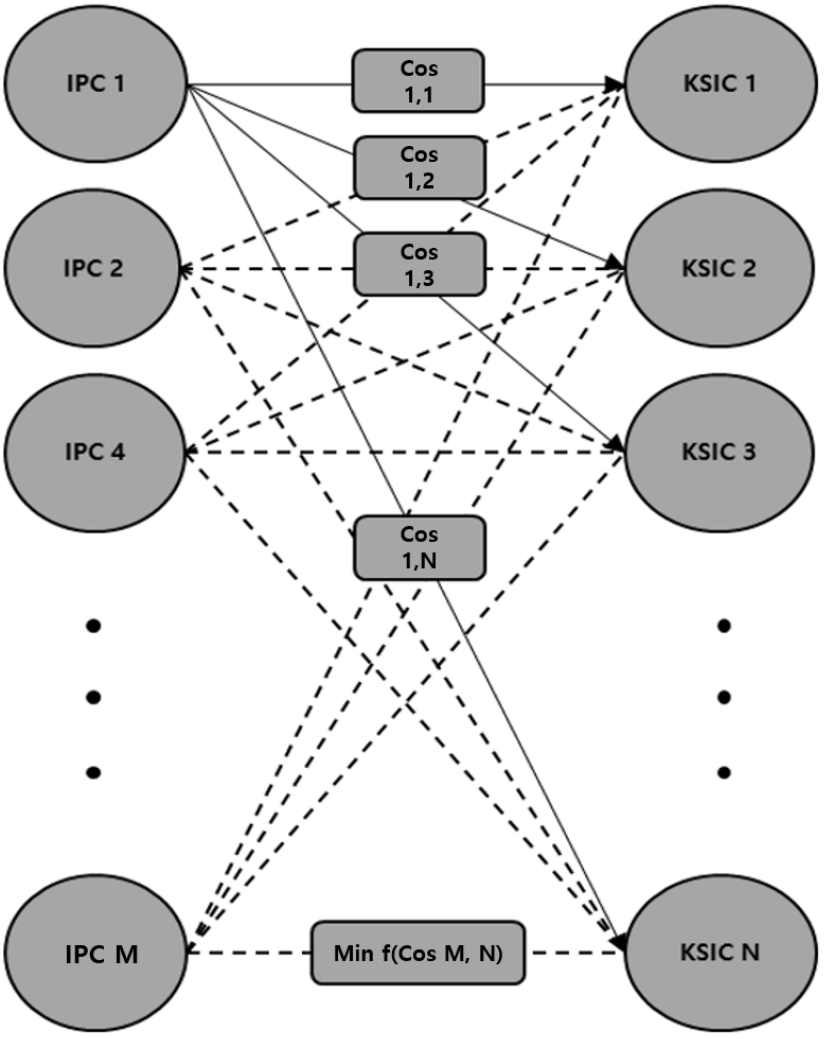
벡터 공간 모델이 최초로 적용된 SMART(System for the Mechanical Analysis and Retrieval of Text) 모델의 표기법에 따라 지정된 가중치 조합을 R에서 수행하기 위한 다양한 계산 기능을 가지고 있는 패키지이다. 따라서 weightSMART에는 TF-IDF를 구하기 위한 다양한 변형식이 있으며 (Luhn, 1957; Spärck Jones, 1972; Manning et al., 2008), 이 중 본 연구에서 사용한 계산식은 아래 <Table 3>과 같다.

단어 빈도(TF, term frequency)에 따른 스키마의 종류에는 4가지의 계산 방법이 있다. Raw count는 문서나 질의 내에서 색인어의 출현 빈도로 중요도를 대신한다. 그리고 Binary는 색인어의 출현 빈도를 무시하고 벡터를 구성하는 색인어에 1의 가중치 부여하는 방법을 말한다. 또한 Log normalization은 색인어의 출현 빈도에 로그 함수 적용하여 이를 보정한 값을 의미하며, Double normalization은 보강된 정규화 출현 빈도를 나타내며 TF를 max TF로 나누어, 그 결과가 0.5~1.0의 값을 갖도록 정규화시키는 계산 방법을 말한다. 한편 문서 빈도(DF, document frequency)를 구하기 위한 여러 가지 계산식 중 본 연구에서는 한 가지 방법만 적용하였다. 구분하자면, Unary는 색인어 출현 빈도만으로 가중치 생성하는 방법을 의미한다. 즉 DF를 고려하지 않는 방식이다. Inverse document frequency는 색인어의 출현 빈도와 역문서 빈도를 곱한 값을 의미하며, 이 때 분자는 전체 문서들의 수이며, 분모는 그 색인어를 포함하고 있는 문서들의 수를 말한다. 그밖에 TF-IDF를 변형한 다양한 계산방법들이 더 있지만, 본 연구에서는 TF(term frequency)의 대표적인 4가지 계산식과 DF(document frequency)의 적용 여부를 조합한 8가지 방식을 적용해보며 가장 정확도가 높은 방법을 선택하였다.

이를 통해 계산된 가중치를 이용하여 KSIC와 IPC 코드별 설명문 간의 유사도를 계산하였다. 구체적으로, <Figure 2>와 같이 각 코드 설명문의 명사 가중치 벡터 간의 Cosine 유사도를 계산하여 문서 간 유사도를 판단하였다. 이는 Cosine 내적 각의 크기가 작을수록 유사도가 높다는 기본적인 가정에 기초한다(Salton et al., 1983). 결과적으로 분류대상문서인 IPC 설명문 별로 가장 유사도가 높은 순서대로 10 순위까지의 KSIC를

〈Table 3〉 TF-IDF transformation calculation

Schema	TF-IDF's weight	Annotation	Equation
TF (term frequency)	raw count	N (natural)	$\mathrm{tf}_{t,d}$
	binary	B (boolean)	$\begin{cases} 1, & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$
	log normalization	L (logarithm)	$1+\log(\mathrm{tf}_{t,d})$
	double normalization	A (augmented)	$0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max(\mathrm{tf}_{t,d})}$
DF (document frequency)	unary	N (no)	1
	Inverse document frequency	T (idf)	$\log \frac{N}{df_t}$



〈Figure 2〉 Similarity calculation process

부여하였다.

그리고 IPC 설명문에 대한 분류 결과의 적절성을 확인하기 위해서 특허청에서 제공한 KSIC-IPC 연계표의 내용과 비교해 보았다. 특허청의 연계표는 주로 KSIC 중분류(2자리 숫자)와 IPC 서브클래스 간의 연결 관계를 나타내고 있다. 본 연구에서는 IPC 서브클래스 코드별 설명문에 대해 유사도 기준 10순위 KSIC 세세분류(5자리) 코드를 매칭하였다. 이를 바탕으로 IPC 코드별로 매칭된 KSIC 세세분류 코드가 <Table 4>와 같은 통계청의 연계표 내용과 얼마나 일치하는 지를 평가하였다.

〈Table 4〉 Part of KSIC-IPC Linkage Table (Korean Intellectual Property Office)

Field	KSIC	IPC
농림어업	A01, A02, A03	A01B27/02, A01C, A01D, A01G, A01H, A01K
광업	B05, B06, B07, B08	C22B, E21D
식료품 제조업	C10	A21D, A23B, A23C, A23D, A23F, A23G, A23J, A23K, A23L, A23P, C12J, C13B, C13K

4. 연구 결과

4.1 결과 해석

수집한 데이터는 각종 정제과정을 거쳤는데, 이 중 가장 많이 출현한 단어는 <Table 5>에 나타낸 바와 같이 '제조'가 20,914개로 제일 많았고, 그 다음으로는 '장치'가 9,847개로 나타났다. 이 둘의 단어 출현 빈도는 상위 20개의 순위로 살펴본 총 단어 출현 빈도의 46%를 차지했다. 이러한 단어들이 많이 출현된 이유는 KSIC에 대분류 C.제조업(10-33)에 해당하는 세세분류 코드가 KSIC 전체 세세분류 코드의 40%를 차지하고 있기 때문이라고 보여진다. 특히 KSIC 설명문에는 기계, 화학, 소재, 금속, 자동차 등의 제조업 내 핵심 산업과 도소매, 서비스업에 관한 키워드가 주로 등장하였다. 반면 IPC 설명문에는 상대적으로 상세한 수준의 키워드들이 포함되어 있는데, 그 결과 결합, 원자, 탄소, 기구, 처리, 부재, 물질 등의 키워드가 상위 출현빈도를 나타내었다.

이처럼 추출된 명사가 가지고 있는 각각의 빈도수와 문서 수를 앞서 <Table 3>에서 설명한 계

〈Table 5〉 Top frequency words list

Rank	Word	TF	DF	Rank	Word	TF	DF
1	제조	20,914	697	11	탄소	1,930	69
2	장치	9,847	648	12	제품	1,819	344
3	도매	3,574	91	13	서비스	1,697	266
4	결합	3,083	359	14	재료	1,667	341
5	기계	2,489	469	15	금속	1,651	265
6	원자	2,157	48	16	자동차	1,599	237
7	제어	2,098	305	17	기구	1,521	328
8	화합물	2,009	104	18	처리	1,494	347
9	소매	1,997	83	19	부재	1,490	180
10	전기	1,950	410	20	물질	1,295	280

산식을 적용하여 가중치를 계산하였다. 그리고 각 문서의 가중치 간 유사도를 계산하였다. 분류대상문서인 IPC 설명문은 Cosine 유사도 값의 비교를 통해 자동적으로 KSIC에 따라 분류될 수 있다. 예컨대 <Table 6>은 LT방식으로 계산한 가중치를 적용하여 KSIC 설명문과 IPC 설명문 간의 유사도를 계산한 결과의 일부를 나타낸다. 첫 번째 열에 있는 IPC A01B 코드는 '농업 또는 임업에 있어서의 토작업; 농기구 또는 기구의 부품, 세부 또는 부속구 일반'이라는 정의를 가지고 있다. <Table 6>에 나타난 15개의 KSIC 코드 중에서 IPC A01B 코드의 설명문과 유사도가 가장 높은 KSIC 코드는 A01110(곡물작물 재배업)인 것을 알 수 있다. 따라서 15개의 KSIC 코드만 고려한다면, IPC A01B의 설명문은 LT의 계산 방식에 따라 KSIC A01110 산업에 해당하는 문서라고 분류할 수 있다.

<Table 7>은 분류대상문서인 IPC A01B에서 A01G까지 5개 코드의 설명문에 대해 유사도가 가장 높은 순서대로 5개까지 KSIC 분류를 수행한 결과를 나타낸다. IPC A01B 내지 A01G 코드는 주로 농업과 농기구에 관한 코드들이다. 이들 코드 설명문에 대한 KSIC 분류 결과 1순위로 C29210(농업 및 임업용 기계제조업)이 주로 등장하였으며, 대체로 농업을 나타내는 A01 계열의 KSIC가 다수 5순위 안에 포함되어 있어 본 연구에서 제안하는 문서의 KSIC 분류 방법이 일정 수준의 정확도를 보이는 것을 확인할 수 있었다.

〈Table 6〉 Example of calculating the similarity between documents

IPC / KSIC	A01B	A01C	A01D	A01F	A01G	A01H	…
A01110	0.0224	0.0059	0.0626	0.0435	0.0822	0.0119	
A01121	0.0109	0.0000	0.0324	0.0124	0.0777	0.0109	
A01122	0.0160	0.0025	0.0272	0.0086	0.1254	0.0292	
A01123	0.0099	0.0450	0.0344	0.0173	0.0909	0.0535	
A01131	0.0103	0.0000	0.0281	0.0171	0.0833	0.0080	
A01132	0.0171	0.0000	0.0316	0.0144	0.0700	0.0000	
A01140	0.0166	0.0055	0.0500	0.0180	0.0823	0.0209	
A01151	0.0000	0.0000	0.0000	0.0000	0.1339	0.0000	
A01152	0.0208	0.0051	0.0415	0.0211	0.1222	0.0198	
A01159	0.0182	0.0041	0.0334	0.0180	0.0958	0.0431	
A01211	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
A01212	0.0002	0.0002	0.0004	0.0000	0.0003	0.0000	
A01220	0.0039	0.0000	0.0000	0.0032	0.0000	0.0000	
A01231	0.0032	0.0000	0.0000	0.0027	0.0000	0.0000	
A01239	0.0002	0.0002	0.0003	0.0000	0.0003	0.0220	
…							

〈Table 7〉 Similarity ranking match results

IPC	Top 1	Top 2	Top 3	Top 4	Top 5
A01B	C29210	G46531	A01411	H49390	C30391
A01C	C29210	C20202	G46732	C20209	G46531
A01D	C29210	A01411	G46531	A01110	A01140
A01F	C29210	A01110	G47211	C29132	H52103
A01G	A01151	A01122	A01152	A02040	A01159

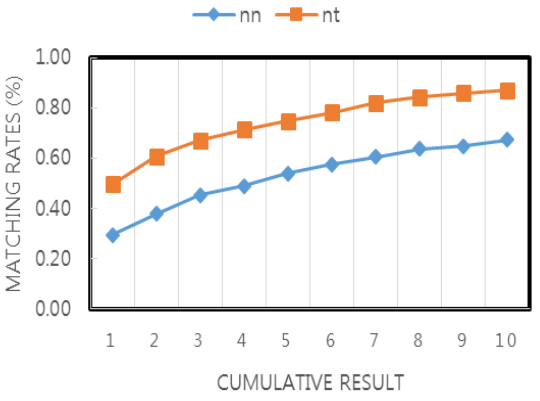
본 연구에서 제안한 방법을 이용하여 각 IPC 코드 별 설명문에 대해 Cosine 유사도 기준 상위 10개의 KSIC 코드를 부여하였다. 분류의 정확도를 보다 객관적으로 판단하기 위하여 통계청에서 제공하는 KSIC-IPC 연계표의 내용과 비교하여, 분류 결과가 얼마나 일치하는지 확인하였다. 하지만 앞서 <Table 3>에 나타낸 바와 같이 문서의 가중치를 설정하는 방법은 매우 다양하기 때문에, 각 방법에 따라 일치도가 달라졌다. <Table 8>은 가중치 설정 방법에 따라 Cosine 유사도 기준 1순위 분류 결과의 일치도와 누적 5순위까지 부여한 분류 결과의 일치도를 나타낸다. 즉 638개 IPC 설명문을 각각 여러 가지 방식으로 문서의 가중치를 부여하고 KSIC 설명문과의 Cosine 유사도를 계산하여, 가장 유사도가 높은 KSIC 코드를 부여한 결과 통계청의 연계표 내용과 얼마나 일치하는지를 나타낸 것이다.

그 결과 LT의 방식으로 문서의 가중치를 부여한 경우, 유사도가 가장 높은 1순위 분류 결과의 일치도가 53%로 다른 방법에 비해 가장 높은 것으로 나타났다. 즉 본 연구의 경우에는 TF(Term Frequency)를 $1+\log tf_{t,d}$, DF는 $\log N/df_t$로 적용하였을 때 가장 우수한 일치도를 보였다. 또한 유사도 상위 5순위까지 선택하였을 때는 일치도가 76%까지 상승하였다. 이는 <Figure 3>에 나타낸 바와 같이, 단순히 단어의 출현 빈도만 TF로 고려한 NN 방법을 적용했을 경우 1순위 분류 결과의 일치도가 29%, 5순위 누적 일치도가 54% 수준인 것에 비해 월등히 우수한 결과이다. 그리고 <Figure 4>는 각 가중치 계산 방식에 따라 유사도 상위 10순위까지 고려하였을 때의 일치도를 나타낸다. 상위 10순위까지 고려하는 경우

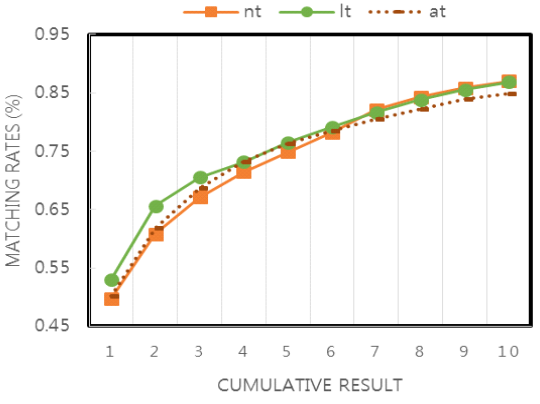
〈Table 8〉 Cumulative match results by weight annotation

DF / TF	N		T	
	1st place cumulative	5st place cumulative	1st place cumulative	5st place cumulative
N	0.29	0.54	0.50	0.75
B	0.37	0.69	0.49	0.76
L	0.43	0.67	0.53	0.76
A	0.39	0.71	0.50	0.76

TF-IDF를 계산하는 대부분의 방식에서 누적 일치도가 88%에 이르는 것으로 나타났다.



〈Figure 3〉 Comparison of matching rates between TF(NN) and TF-IDF(NT)



〈Figure 4〉 Comparison of matching rates between TF-IDF

4.2 토의

본 연구에서 제안한 문서의 KSIC 분류 결과를 통계청에서 제공한 연계표와 비교했을 때의 일치도는 가중치 계산 방법에 따라 차이가 있었다. 먼저 <Figure 3>을 통해 TF 보다는 TF-IDF가 더 우수한 일치도를 나타냈는데, 구간별로 20%~22%의 차이가 있었다. 이러한 차이가 나는 이유는 TF의 경우에 상대적으로 문서의 길이가 길수록 많은 정보량을 갖게 되는 왜곡현상으로 인해 모델 전반의 분류 일치도를 저하시키기 때문이다. 이러한 현상은 여러 선행연구들을 통해서도 밝혀진 바 있다. Lee and Kim(2009)은 특정분야의 뉴스문서 집합에서 키워드를 추출하는데 TF 값의 편중 현상을 방지하기 위해 IDF 인자를 도입한 TF-IDF를 변형하여 6개의 모델의 성능을 비교한 바 있다. Noh et al.(2017)는 TF에 주어진 단어가 문서 내에서 많이 출현할수록 상대적으로 더 중요하다는 가정이 반영되어 있지만, IDF를 이용해 문서 내에서 주요 의미를 가지는 단어를 분별할 수 있다고 보고하였다. 그리고 이렇게 구한 TF-IDF에 시간 속성을 더하여 SNS의 핫 토픽을 예측하는 기법을 제안한 바 있다.

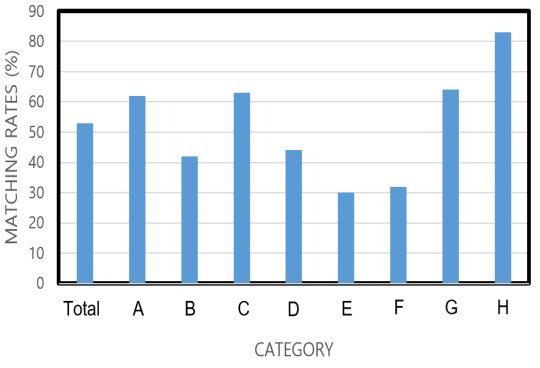
한편 TF-IDF를 계산하는 방식에 따라서도 결과에 큰 차이가 있었다. 가장 간단하게 단어의 출현 빈도만 고려하는 NT 방식에 비해 AT와 LT의 순서대로 일치도가 증가하였다. 그 이유는 TF를 보정하여 IDF와 계산함으로써 보다 유의미한 정보량을 할당할 수 있기 때문이다. 이와 관련해서 Lee and Kim(2009)은 TF-IDF를 계산하기 앞서 TF를 BTF(Basic Term Frequency)와 NTF(Normalized Term Frequency) 두가지로 구분하였다. 그 결과 (TF) x IDF와 (1+log(TF)) x IDF로 만든 조합 식들이 TF-IDF의 가장 기본형에 해당하는 BTF-IDF 보다 개선된 성능을 보인다고 보고하였는데, 이는 본 연구 결과와도 유사하다. 본 연구의 결과에 따르면, 가장 성능이 좋은 LT 방식으로 가중치를 계산하여 유사도 1순위만 고려하는 경우, 통계청의 연계도와 비교한 일치도는 53%로 나타났다. 이에 대해 더 높은 일치도 결과

를 얻어내지 못한 것에 대해서는 크게 3가지 이유를 생각해 볼 수 있다.

첫째, 일반적으로 하나의 IPC 코드가 하나의 세부 산업에만 관련되는 것이 아니기 때문이다. 산업을 분류하는 KSIC와 기술군을 분류하는 IPC는 태생적으로 다른 목적, 범위, 분류 기준을 바탕으로 분류되었으며 서로를 고려하여 분류체계가 마련된 것이 아니다. 따라서 KSIC와 IPC는 서로 1-1 혹은 1-N(多)의 매칭이 되기 보다는 N(多)- N(多) 매칭이 이루어질 것을 예상할 수 있다. 하지만 통계청에서 작성한 KSIC-IPC 연계표는 하나의 IPC 코드가 하나의 KSIC로 1-N 분류가 되어있다. 따라서 본 모델의 분류 결과로 유사도 1순위만 고려했을 경우 100%의 일치도를 보이는 것은 기대하기 어려우며, 후순위의 결과까지 동시에 고려할 필요가 있음을 알 수 있다. 예를 들어 <Table 7>에 나와있는 IPC 중 '쟁기, 써레, 농작업기' 등을 나타내는 A01B와 '파종기, 이식기, 퇴비살포기, 비료살포기' 등을 나타내는 A01C, '예취기, 갈퀴, 굴취기, 수확기' 등을 나타내는 A01D, 그리고 '탈곡기, 짐꾸리기 압축기, 짚 운반장치' 등을 나타내는 A01F는 '농업 및 임업용 기계 제조업'을 나타내는 1순위 KSIC 분류 결과인 C29210으로 알맞게 매칭이 된 것으로 보여진다. 하지만, 통계청에서 고시한 KSIC-IPC 연계표에는 KSIC C29(C2920) 중분류에는 A01B와 A01F만 매칭이 되어 있다. 따라서 해당 연계표로 검증 작업을 수행한 본 모델에서 다소 알맞게 분류한 것으로 보여지는 A01C와 A01D는 오분류로 평가되기 때문에 이러한 현상은 본 모델을 검증하는 단계에서 일치도를 저하시킨다. 더욱이 KSIC-IPC 연계표의 KSIC C29(C2920) 중분류와 매칭되어 있는 A01B와 A01F의 2순위 KSIC 분류 결과는 각각 G46531과 A01110로, '농업용 기계 및 장비 도매업'과 '곡물 및 기타 식량작물 재배업'을 나타낸다. 즉, A01B와 A01F는 본 모델의 후순위 결과에도 관련이 있으므로, KSIC와 IPC는 앞서 말했듯이 N(多)- N(多)으로 매칭되는 것이 보다 합리적인 것으로 보여진다. 따라서 이러한 연계표와의 일치도 검증은 완벽한 결과를 얻는데 제약을 준다.

둘째, KSIC 설명문과 IPC 설명문에서 사용하는 용어에 차이가 존재하기 때문이다. 가령 전기를 나타내는 H 기술군의 경우 사용되는 용어가 서로 잘 일치하는 반면, '고정구조물'과 '기계공학; 조명; 가열; 무기; 폭파'를 나타내는 E, F 기술군은 그렇지 못하다. 예를 들어 IPC 중 '저항기(H01C)', '변성기(H01F)', '콘덴서(H01G)', '진공관(H01J)', '인쇄회로(H05K)' 등 '전기'를 나타내는 H 기술군의 경우 사용되는 용어가 각각 KSIC 중 '전자저항기 제조업(C26293)', '전자코일, 변성기 및 기타 전자유도자 제조업(C26295)', '전자축전기 제조업(C26292)', '전자관 제조업(C26292)', '인쇄회로기판 제조업(C26221)' 등에서 사용하는 용어와 비교적 잘 일치하였다. 반면 통계청의 연계표에 따르면 IPC E01B(궤도; 궤도용 공구, 모든 종류의 철도건설용 기계)는 KSIC C312(철도장비 제조업)에 연결되어 있다. 그런데 E01B 설명문에 가장 많이 등장한 '레일'이라는 용어는 KSIC에서는 대부분 '궤도'로 표현되어 있다. 또한 IPC E02B(수공; 水工) 설명문에서 가장 많이 등장하는 '게이트(gate)'라는 용어는 KSIC에서는 '수문'이나 '댐'으로 표현되고 있다. 이 밖에도 E, F 기술군의 IPC 설명문에는 영어식 표현이 그대로 사용되어 KSIC 설명문과의 용어 일치도를 낮추는데, 예를 들면 건축과 관련된 '버켓', '시스턴', '슬릿' 등의 용어와 무기 제조와 관련된 '브리취', '카트리지' 등을 용어를 꼽

을 수 있다. 그 결과, <Figure 5>에 나타낸 바와 같이 IPC 기술군 별 일치도에 큰 차이가 있어 H 기술군의 유사도 1순위 일치도가 83%에 육박하였으나, E 기술군의 경우 30% 수준에 그쳤다. 즉 분야별 사용된 용어의 일치 정도가 영향을 미친 것으로 판단된다.



〈Figure 5〉 Comparison of matching rates between categories

셋째, 본 연구에서 활용한 코드 설명문이 해당 코드를 충분히 설명하는 내용과 분량의 말뭉치로서 부족한 경우가 존재한다. KSIC 설명문의 경우 색인어 정보를 함께 포함하고 있기 때문에 어느 정도 문서의 양을 확보할 수가 있다. 하지만 IPC의 경우에는 몇몇 코드가 예컨대 '수확; 예취', '탈곡' 등으로 내용을 설명하기에 매우 짧아 정확한 분류에 어려움을 주기도 한다. 이러한 한계를 감안한다면 본 연구에서 얻어진 일치도 결과는 정성적 분류 방법에 비교할 만한 수준으로 판단된다. 또한 다양한 디지털 정보를 KSIC를 기준으로 분류하여 중소기업에게 맞춤형 정보를 제공하는데 충분한 활용 가치가 있는 것으로 판단된다.

5. 결론

본 연구에서는 기업의 기술사업화 의사결정에 도움이 될 수 있는 맞춤형 정보를 자동으로 분류하여 제공하기 위하여, 기업의 사업 성격을 나타내는 KSIC를 기준으로 정보를 분류하고자 하였다. 이를 위해 KSIC 각 코드별 설명문을 수집하고 벡터 공간 모델을 이용하여 분류 대상 문서와의 유사도를 계산하여 가장 적합한 KSIC 코드를 제시하는 모델을 제시하였다. 그리고 IPC 데이터를 수집한 후 KSIC를 기준으로 분류하고, 이를 특허청에서 제공하는 KSIC-IPC 연계표와 비교함으로써 본 방법론을 검증하였다.

검증 결과 TF-IDF 계산식의 일종인 LT 방식을 적용하였을 때 가장 높은 일치도를 보였는데, IPC 설명문에 대해 1순위 매칭 KSIC의 일치도는 53%, 5순위까지의 누적 일치도는 76%를 보였다. 이는 기존 연계표의 한계와 양 분류체계에서 사용하는 용어의 차이 등을 감안하면 충분히 높은 수준으로 판단된다. 이를 통해 보다 정량적이고 객관적으로 중소기업이 필요로 할 기술, 산업, 시장 정보에 대한 KSIC 분류 작업이 가능하다는 점을 확인할 수 있었다.

본 연구의 결과는 이종 분류 체계간 연계를 필요로 하는 다양한 학분 분야에도 활용될 수 있다. 경제·사회 현상을 설명하는 다양한 통계자료는 서로 다른 분류체계에 따라 작성되고 있는데, 보다 복잡한 현상을 이해하고 예측하기 위해서는 이종 간 연계데이터에 대한 이해와 학습이 요구된다 (Yoo et al., 2015). 이 때 이종 분류 체계간 연계가 요구되는데. 특히 새로운 분류체계와의 연계 문제에 있어서 전문가의 정성적 판단이 어렵거나 충분한 학습 데이터가 부족한 경우가 많으며, 이것이 현재 많이 사용되고 있는 기계학

습 기반 분류 방식을 활용하기 어렵게 한다. 본 연구에서 제안한 이종 분류 체계의 설명문 간 유사도를 기준으로 분류 체계를 연결하는 방법은 매우 간단하면서도 다양한 학문 분야에서 실질적으로 이러한 문제를 해소할 수 있는 실마리를 제공할 것으로 기대된다.

따라서 본 연구는 현실세계의 문제를 해결하는데 있어서 최근 각광받고 있는 지도 학습 기법이 우수한 예측력에도 불구하고 태생적으로 가질 수 밖에 없는 한계에 대해 범용적인 활용 측면에서 문제 제시를 하는 바이며, 이를 비지도 학습 기법인 벡터 공간 모델 기법을 사용하여 기술사업화 목적의 맞춤형 문서 자동 분류 사례를 통해 다양한 가능성을 시사하고 있다.

참고문헌(References)

Aha, D. W., D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning,* Vol.6, No.1(1991), 37~66.

Beel, J., B. Gipp, S. Langer, and C. Breitinger, "paper recommender systems: a literature survey," *International Journal on Digital Libraries*, Vol.17, No.4(2016), 305~338.

Byun, S., Lee, D., and Kim, N,. “Methodology for Identifying Issues of User Reviews from the Perspective of Evaluation Criteria: Focus on a Hotel Information Site,” *Journal of Intelligence and Information Systems,* Vol.22, No.3(2016), 23~43.

Chang, J., "Using the MeSH Hierarchy to Index Bioinformatics Articles," *CS224N/Ling237 Final Projects,* (2000), 1~10.

Chang, J. Y., "A Study on Research Trends of Graph-Based Text Representations for Text Mining," *The Journal of The Institute of Internet, Broadcasting and Communication*, Vol.13, No.5(2013), 37~47.

Choi, H. B., "An Artificial Neural Network for Local Library’s Book Recommender System," *Journal of Korean Institute of Information Technology*, Vol.14, No.9(2016), 109~118.

Cleverdon, C., "Optimizing Convenient Online Access to Bibliographic Databases," *Information Services and Use*, Vol.4, No.12 (1983), 37~47.

Cooper, W. S. "Getting beyond boole," *Information Processing & Management,* Vol.24, No.3(1988), 243~248.

Craven, M., et al., *Learning to extract symbolic knowledge from the World Wide Web*, Carnegie-mellon univ pittsburgh pa school of computer Science, 1998.

Craven, M., et al. "Learning to construct knowledge bases from the World Wide Web," *Artificial intelligence,* Vol.118, No.1 (2000), 69~113.

Dillon, M. "Introduction to modern information retrieval: G. Salton and M. McGill", McGraw-Hill, New York, 1983.

Drucker, P., *Post-capitalist society*, Routledge, 2012.

Gudivada, V. N., V. V. Raghavan, W. I. Grosky, and R. Kasanagottu, "Information retrieval on the world wide web", *IEEE Internet Computing*, Vol.1, No.5(1997), 58~68.

Guide to the International Patent Classification, WIPO, 2017.

Hamedani, M. R., and S. W. Kim, "A Comparative Study of Vector Space and Probabilistic

Models in Computing Similarity of Scientific Papers," *Communications of the Korean Institute of Information Scientists and Engineers*, Vol.20, No.3(2014), 186~190.

Hong, J. S., Kim, N., and Lee, S., "A Methodology for Automatic Multi-Categorization of Single-Categorized Documents," *Journal of Intelligence and Information Systems,* Vol.20, No.3(2014), 77~92.

Jeon, H. C., and J. M. Choi, "PIRS : Personalized Information Retrieval System using Adaptive User Profiling and Real-time Filtering for Search Results," *Journal of Intelligence and Information Systems*, Vol.16, No.4(2010), 21~41.

Jeong, Y. M., "Information Retrieval Theory", Gumi Trade Publishing Department, 1993.

Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *European Conference on Machine Learning(ECML)*, 1988.

Kim, D. and Yu, S. J., "Reliability Analysis of VOC Data for Opinion Mining," *Journal of Intelligence and Information Systems,* Vol.22, No.4(2016), 217~245.

Kim, G., "Data Mining for Spam Email Classification," *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol.6, No.7 (2016), 37~47.

Kim, H. J., and J. Y. Chang, "A Semantic Text Model with Wikipedia-based Concept Space," *The Journal of Society for e-Business Studies*, Vol.19, No.3(2014), 107~123.

Kim, S. I., and H. S. Kim, "An Automatic Web Page Classification System Using Meta-Tag," *The Korean Institute of Communications and Informaion Sciences,* Vol.38, No.4(2013), 291~297.

Korea Standard Industry Classification(KSIC) 9th Amendment, Statistics Korea, 2007.

Lang, K., "Newsweeder: Learning to filter netnews," *Machine Learning Proceedings 1995*, (1995), 331~339.

Lee, H. K., S. Yang, and Y. J. Ko, "Feature Expansion based on LDA Word Distribution for Performance Improvement of Informal Document Classification," *Korea Institute of Information Scientists and Engineers*, Vol.43, No.9(2016), 1008~1014.

Lee, J. M., "UN's Sustainable Development Goals (SDGs) Oriented Research Trend in publications of Korean Society of Rural Planning, 1995-2016: quantitatively analyzed with the Vector Space Model," *Journal of Korean Society of Rural Planning*, Vol.23, No.2(2017), 29~42.

Lee, J. H., M. H. Kim, and Y. J. Lee, "Ranking documents in thesaurus-based Boolean retrieval systems," *Information Processing & Management,* Vol.30 No.1(1994), 79~91.

Lee, S., and H. J. Kim, "Keyword Extraction from News Corpus using Modified TF-IDF," *The Journal of Society for e-Business Studies*, Vol.14, No.4(2009), 59~73.

Lee, S., G. Lee, O. Hwang, and S. Noh, "Developing Movie Recommendation System Reflecting Movie Viewers' Preferences," Journal of Intelligence and Information Systems 2007 Fall Conference, (2007), 507~513.

Lewis, D. D., and W. A. Gale, "A sequential algorithm for training text classifiers,"

Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Springer-Verlag New York, Inc., 1994.

Lewis, D. D., and K. A. Knowles, "Threading electronic mail: A preliminary study," *Information processing & management,* Vol.33, No.2(1997), 209~217.

Luhn, H. P., "A statistical approach to mechanized encoding and searching of literary information" *IBM Journal of research and development,* Vol.1, No.4(1957), 309~317.

Manning, C. D., P. Raghavan, and H. Schtze, "Document and query weighting schemes," *Introduction to Information Retrieval,* (2008), 128.

Ministry of SMEs and Startups, "Status of SMEs in Korea", 2014.

Mooney, R. J., and L. Roy, "Content-based book recommending using learning for text categorization," *Proceedings of the fifth ACM conference on Digital libraries*, ACM, (2000).

National Information Society Agency, "2016 The Report on the Digital Divide", 2016.

Noh, Y., J. Lim, K. Bok, J. Yoo, “Hot Topic Prediction Scheme Using Modified TF-IDF in Social Network Environments,” *KIISE Transactions on Computing Practices*, Vol.23, No.4(2017), 217~225.

Park, C. H., S. S. Youm, and J. M. Lee, "The Effect of User-Centered Categorization System of Homepages on Directory Search," *Korean Journal of Cognitive Science*, Vol.11, No.1(2000), 47~65.

Pazzani, M. J., J. Muramatsu, and D. Billsus, "Syskill & Webert: Identifying interesting web sites," *AAAI/IAAI, Vol. 1*. 1996.

Ponte, J. M., and W. B. Croft, "A language modeling approach to information retrieval," *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, (1998).

Radecki, T. "Trends in research on information retrieval—the potential for improvements in conventional boolean retrieval systems," *Information Processing & Management,* Vol.24, No.3(1988), 219~227.

Ruiz, M. E., and P. Srinivasan, "Hierarchical text categorization using neural networks," *Information Retrieval,* Vol.5, No.1(2002), 87~118.

Salton, G., A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM,* Vol.18, No.11(1975), 613~620.

Salton, G., "Historical Note: The Past Thirty Years in Information Retrieval," *Jounal of the American Society for Information Science*, Vol.38, No.5(1987).

Salton, G. "Automatic text processing: The transformation, analysis, and retrieval of," *Reading: Addison-Wesley,* (1989).

Sebastiani, F., "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, Vol.34, No.1(2002), 1~47.

Shavlik, J., and T. Eliassi-Rad, "Intelligent agents for web-based tasks: An advice-taking approach," *AAAI/ICML Workshop on Learning for Text Categorization*, 1998.

Sparck Jones, K., "A statistical interpretation of

term specificity and its application in retrieval," *Journal of documentation,* Vol.28, No.1(1972), 11~21.

Vapnik, V. *Statistical learning theory, 1998*, Wiley, New York, 1998.

Witten, I. H., A. Moffat, and T. C. Bell, *Managing gigabytes: compressing and indexing documents and images*, Morgan Kaufmann, 1999.

Yang, Y. "Expert network: Effective and efficient learning from human decisions in text categorization and retrieval," *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag, New York, Inc., 1994.

Yang, Y., and J. O. Pedersen. "A comparative study on feature selection in text categorization," *Icml*, Vol. 97, (1997).

Yang, Y. and X. Liu, "A Re-examination of Text Categorization Methods," *Proceedings of the 22h Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 99)*, (1999), 42~49.

Yang, Y. "An Evaluation of Statistical Approaches to Text Categorization," *Journal of Information Retrieval*, Vol.1, No.1(1999), 67~88.

Yoo, H. S, J. H. Seo, S.-P. Jun, J. Seo, “A Study on an Estimation Method of Domestic Market Size by Using the Standard Statistical Classfications,” *Journal of Korea Technology Innovation Society*, Vol. 18, No. 3(2015), 387~415.

Abstract

A Study on Automatic Classification Model of Documents Based on Korean Standard Industrial Classification

Jae-Seong Lee* · Seung-Pyo Jun** · Hyoung Sun Yoo***

As we enter the knowledge society, the importance of information as a new form of capital is being emphasized. The importance of information classification is also increasing for efficient management of digital information produced exponentially. In this study, we tried to automatically classify and provide tailored information that can help companies decide to make technology commercialization. Therefore, we propose a method to classify information based on Korea Standard Industry Classification (KSIC), which indicates the business characteristics of enterprises. The classification of information or documents has been largely based on machine learning, but there is not enough training data categorized on the basis of KSIC. Therefore, this study applied the method of calculating similarity between documents. Specifically, a method and a model for presenting the most appropriate KSIC code are proposed by collecting explanatory texts of each code of KSIC and calculating the similarity with the classification object document using the vector space model. The IPC data were collected and classified by KSIC. And then verified the methodology by comparing it with the KSIC-IPC concordance table provided by the Korean Intellectual Property Office. As a result of the verification, the highest agreement was obtained when the LT method, which is a kind of TF-IDF calculation formula, was applied. At this time, the degree of match of the first rank matching KSIC was 53% and the cumulative match of the fifth ranking was 76%. Through this, it can be confirmed that KSIC classification of technology, industry, and market information that SMEs need more quantitatively and objectively is possible. In addition, it is considered that the methods and results provided in this study can be used as a basic data to help the qualitative judgment of experts in creating a linkage table between heterogeneous classification systems.

* University of Science & Technology
** Div. of Data Analysis, Korea Institute of Science & Technology Information/ University of Science & Technology
*** Corresponding Author: Hyoung Sun Yoo
Div. of Data Analysis, Korea Institute of Science & Technology Information/ University of Science & Technology
66 Hoegi-ro, Dongdaemun-gu, Seoul 02456, Korea
Tel: +82-2-3299-6173, Fax: +82-2-3299-6139, E-mail: hsyoo@kisti.re.kr

Key Words : Automatic Document Classification, Korea Standard Industry Classification, Text mining, Vector space model, Natural language processing

Received : May 31, 2018 Revised : July 9, 2018 Accepted : August 20, 2018
Publication Type : Regular Paper(Fast-track) Corresponding Author : Hyoung Sun Yoo

저 자 소 개

이 재 성
공주대학교 경영학 및 독문학 학사를 하며 4차 산업혁명에 관심을 갖게 되었다. 현재는 과학기술연합대학원대학교 과학기술경영정책학과에서 통합과정 중이며 한국과학기술정보연구원 데이터분석본부에서 학생연구원으로 재직 중에 있다. 주요 관심 연구분야로는 데이터 마이닝과 기계학습 및 인공지능 기법을 활용한 정부의 중소기업 기술사업화 지원정책과 국가 과학기술 연구개발 관리에 관심을 가지고 있다.

전 승 표
KAIST에서 경영학으로 석사학위를 취득하고, 고려대학교에서 과학관리학 전공으로 이학박사를 취득했다. 현재 한국과학기술정보연구원 데이터분석본부에 책임연구원으로 재직 중이며, 과학기술연합대학원대학교 과학기술정책학과 부교수로 재직 중이다. Technological forecasting and social change, Scientometrics, Energy policy, Internet research 등 해외학술지와 한국기술혁신학회지, 지능정보연구 등 국내학술지에 주저자로 다수의 논문을 게재했다. 주요 관심분야는 빅데이터를 활용한 수요 예측, 유망 기술 탐색, 기술 가치평가, 산업시장분석 등을 위한 지능형 정보 시스템 연구이다.

유 형 선
한국과학기술원에서 공학 박사학위를 취득하고 현재 한국과학기술정보연구원 데이터분석본부 책임연구원으로 재직 중이다. 과학기술연합대학원대학교 과학기술경영정책학과의 부교수를 겸임하고 있다. 관심 연구분야는 산업시장분석 방법론, 중소기업 R&D 정책, 행위자 기반 모델링, 복잡계 네트워크 등이다.