디지털산업정보학회 논문지 제15권 제4호-2019년 12월
http://dx.doi.org/10.17662/ksdim.2019.15.4.103

효율적인 특허정보 조사를 위한 분류 모형

김 영 호** · 박 상 성*** · 장 동 식****

A Novel Classification Model for Efficient Patent Information Research

Kim Youngho · Park Sangsung · Jang Dongsik

〈Abstract〉

A patent contains detailed information of the developed technology and is published to the public. Thus, patents can be used to overcome the limitations of traditional technology trend research and prediction techniques. Recently, due to the advantages of patented analytical methodology, IP R&D is carried out worldwide. The patent is big data and has a huge amount, various domains, and structured and unstructured data characteristics. For this reason, there are many difficulties in collecting and researching patent information. Patent research generally writes the Search formula to collect patent documents from DB. The collected patent documents contain some noise patents that are irrelevant to the purpose of analysis, so they are removed. However, eliminating noise patents is a manual task of reading and classifying technology, which is time consuming and expensive. In this study, we propose a model that automatically classifies The Noise patent for efficient patent information research. The proposed method performs Patent Embedding using Word2Vec and generates Noise seed label. In addition, noise patent classification is performed using the Random forest. The experimental data is published and registered with the USPTO among the patents related to Ocean Surveillance & Tracking Network technology. As a result of experimenting with the proposed model, it showed 73% accuracy with the label actually given by experts.

Key Words :Patent Information Research, Noise Patent Classification, Word2Vec, Random Fores

Ⅰ. 서론

모든 사물이 초연결되는 4차 산업혁명 시대의 경쟁력은 특허이다. 특허는 개발된 기술의 상세한 정보를 포함하고 있으며 대중에 공개된다. 따라서 이를 활용하여, 기술의 융복합 및 새로운 분야 도출이 가능하다. 특허는 기존의 기술동향조사 및 예측기법의 한계를 극복하고 있다. 왜냐하면, 특허문서 자체가 기술을 모두 포함하고 있으므로 해당 문서에 관한 분석은 곧, 기술을 분석하는 것과 같은 효과가 있기 때문이다[1-2].

최근에는 특허를 활용한 분석방법론의 이점 때문

* 이 논문은 2019년 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임.(한국연구재단-NRF-2017R1A2B1010208).
** 고려대학교 산업경영공학부 석 · 박사통합과정(제1저자)
*** 청주대학교 빅데이터통계학과 조교수(교신저자)
**** 고려대학교 산업경영공학부 교수

에 전 세계적으로 IP R&D(Intellectual Property Research and Development)를 진행하고 있다. IP R&D는 지식재산권 중심의 연구개발 방법으로서, 먼저 관련 특허 포트폴리오를 분석한 후에 R&D를 수행하는 것이다. IP R&D는 경쟁자 및 선행 특허들을 사전에 분석하므로 중복연구 회피, 핵심·원천특허 등을 조기에 선점할 수 있는 이점이 존재한다. 이러한 장점에도 불구하고 특허는 빅데이터로서, 매년 전 세계에서 방대한 양이 출원 및 등록된다. 또한, 다양한 기술분야를 다루고 정형·비정형 데이터의 특성을 모두 포함하고 있다[3]. 이러한 특허빅데이터의 (Patent bigdata) 특성 때문에 특허 정보 수집 및 조사에는 많은 어려움이 존재한다.

특허 조사는 일반적으로 특정 기술분야의 검색식을 작성하고 DB로부터 특허를 검색 및 수집한다. 다음으로, 수집된 특허는 분석 목적과 무관한 노이즈(Noise) 특허를 다소 포함하고 있으므로, 노이즈 특허 제거 작업을 수행한다[4]. 그러나 노이즈 특허 제거 작업은 대부분 직접 기술을 읽고 분류하는 수작업으로 이루어진다. 이는 많은 시간과 비용을 소모하고 신기술 분야의 적절한 전문가를 모색하기 어려운 단점이 존재한다. 또한, 분석자에 따라 노이즈 분류 기준이 상이하다. 최근에는 이러한 문제점을 해결하기 위해서 시간을 단축시키고 정확도를 높이는 방법론에 대한 요구가 급증하고 있다.

본 논문에서는 효율적인 특허정보 조사를 위한 노이즈 특허 분류 모형을 제안한다. 제안하는 방법은 Word2Vec을 활용하여 기술적인 내용을 고려한 특허 임베딩(Patent embedding)을 수행한다. 또한, 거리기반 가중치 행렬을 구축하여 노이즈 특허 분류를 위한 Seed label을 생성한다. 마지막으로, 이를 Bootstrap 기반의 Ensemble 모형인 Random forest로 학습 및 분류한다.

Ⅱ. 관련연구

2.1 IP R&D

IP R&D는 지식재산권 및 관련 정보들을 활용하여 R&D를 기획하고 기업경영에 전략적으로 활용하는 것이다[5]. 애플과 삼성의 특허전쟁으로 인하여 선 연구개발 후 권리확보의 문제점이 대두되었으며, 연구개발 트렌드가 IP R&D로 변화하고 있다. IP R&D의 장점은 기존의 특허 포트폴리오을 분석함으로써, 중복 연구의 방지 및 보다 강력한 특허 포트폴리오 구축이 가능하다. 또한, 경쟁기업들로부터 진입장벽을 높임과 동시에 기술 라이센싱을 가능하게 하여, 자사의 경쟁력을 높일 수 있다. 더 나아가 기술의 표준화 선도가 가능한 표준 특허를 창출함으로써, 관련 업계의 기술 우위와 수익 창출을 도모할 수 있다.

성공적인 IP R&D를 위해서는 효율·효과적인 특허 정보조사가 수반되어야 한다. 그러나 특허는 빅데이터로서 조사에 다양한 문제점이 존재한다. 그 중 노이즈 특허 문서의 분류는 시간 및 비용이 많이 소모되는 작업이다. 본 논문에서는 이를 해결하고자, 노이즈 특허를 자동으로 분류하는 모형을 제안한다.

2.2 Word2Vec

Word2Vec은 Mikolov et al. [6]이 제안한 단어 분산 표상 기법이다. 이는 같은 맥락(Context)을 가진 단어들은 Embedding 공간상에서 근접하게 위치한다는 분포 가설(Distributional hypothesis)을 전제로 한다[7, 8]. Word2Vec은 신경망 기반으로, 예측 단어 구분에 따라 CBOW(Continuous Bag of Words)와 Skip-gram 방식으로 나뉜다. 두 방식 모두 유사한 단어끼리는 근접하도록 학습하여 벡터값을 부여한다. 본 논문에서는 일반적으로 성능이 우수하다고 알려진 Skip-gram을 사용한다. 아래의 <식 1>은

Skip-gram의 목적 함수이다.

$$p(c|w) = \frac{\exp(x_w^T v_c)}{\sum_{k=1}^{K} \exp(x_w^T v_k)} \quad \text{<식 1>}$$

<식 1>에서 c는 주변단어, w는 중심단어를 나타낸다. Skip-gram은 <식 1>이 최대 값을 가지도록 학습된다.

기존에 다양한 연구들이 문서 분류에 Word2Vec을 활용하였다[9-12]. Kim et al. [9]은 Word2Vec을 사용하여 문서 검색에 활용 가능한 특징벡터 도출을 연구하였다. 이들은 대상문서에서 Word2Vec으로 단어 유사도 행렬을 구축하고, 질의문서(Query documents)로부터 도출한 단어별 중심점수 벡터와 내적 하였다. Seo et al. [10]은 감성분석을 위해 Word2Vec으로 단어 유사도 행렬을 구축하고 그래프 기반 준지도 학습(Semi-Supervised learning)으로 분류하였다. 이들은 준지도 학습 적용 시에 분류된 Label이 존재하는 데이터를 사용하였다. Kim and Park [11]는 특허의 IPC 분류를 위해 Word2Vec으로 가중치 행렬을 생성하고 이를 양방향 장단기 기억 네트워크로 학습 및 분류하였다. 이와 같은 선행연구들은 Word2Vec을 사용하여 단어 가중치의 생성 및 문서 분류를 수행하였다. 그러나 질의문서 또는 기존에 분류된 Label이 필요하므로, 학습을 위한 Label이 없는 노이즈 특허분류에 적용이 어렵다.

Jeong et al. [12]은 휴대폰 리뷰의 분류를 위해 Word2Vec을 활용하였다. 이들은 먼저 가중치 행렬을 구축하고, 그래프 기반 방법론을 사용하여 중심단어를 산출하였다. 또한, 중심단어들만을 이용하여 기능별로 휴대폰 리뷰를 분류하였다. 그러나 중심단어만을 사용하여 분류를 진행한다면, 데이터의 정보손실 가능성이 존재한다. 본 연구에서는 이와 같은 문제점을 해결하기 위해 초기에 부여한 Seed label의 전체 단어들을 Bootstrap 기반의 Random forest로 학습한다.

2.3 Random forest

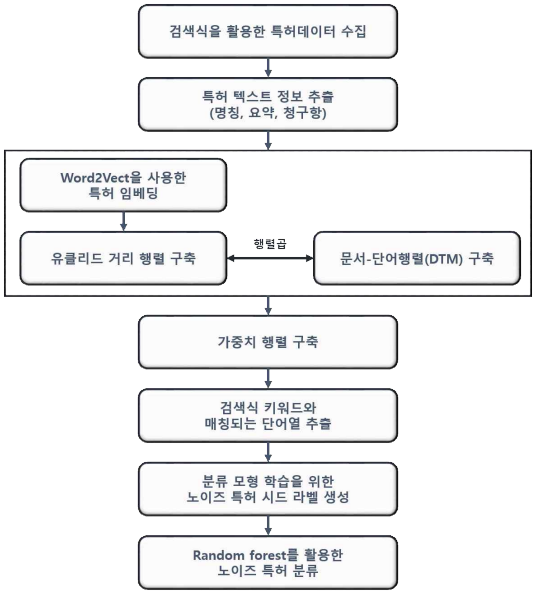
Random forest는 Tree 기반의 Ensemble 분류 모형이며, 일반적으로 좋은 성능을 보인다. 이는 주어진 데이터에서 n개의 자료를 이용한 Bootstrap 표본을 생성한다. 또한, 입력 변수들을 무작위로 추출하고 서로 다른 Decision Tree를 생성한 후, 선형결합하여 최종 분류 모형을 구축한다[13, 14]. Random forest는 회귀(Regression)와 분류(Classification) 문제에 모두 적용이 가능하다[15]. 일반적으로 분류 문제에서의 선형결합 방식은 다수결 원칙(Majority voting)을 적용한다.

기존에 다양한 연구들이 문서 분류에 Random forest를 활용하였다[16, 17]. W. T. Aung et al. [16] 웹페이지 문서들의 카테고리 다중 분류를 위해 Random forest를 사용하였다. A. Onan et al. [17]은 다양한 문서 데이터에 키워드 추출과 Ensemble 알고리즘들을 적용하여 분류 성능을 비교하였다. 그중 Random forest는 약 93.8%의 일치율을 보이는 것으로 가장 우수하였다. 본 논문에서는 노이즈 특허 Seed label을 Random forest로 학습 및 분류한다.

Ⅲ. 연구 방법

3.1 연구 모형

본 연구에서는 효율적인 특허정보 조사를 위하여 노이즈 특허를 자동으로 분류하는 모형을 제안한다. 제안하는 모형은 아래 <그림 1>과 같이 수행된다.



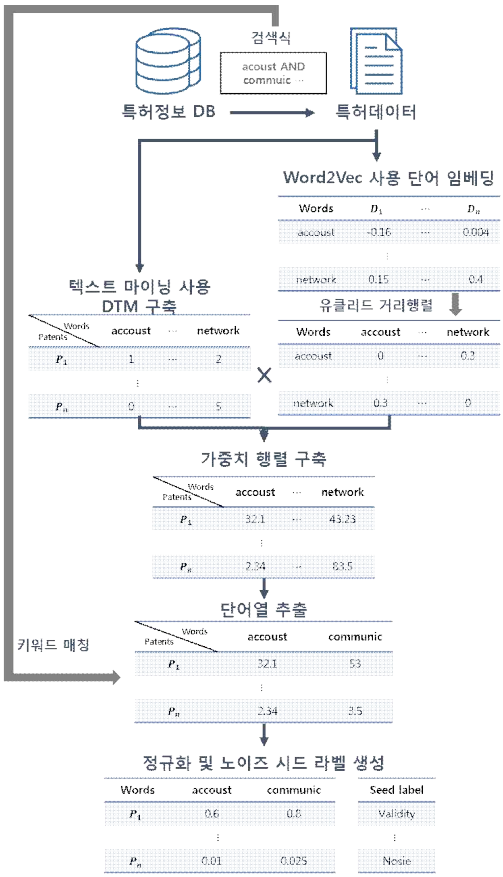
<그림 1> 제안하는 연구 모형

<그림 1>의 연구 모형 단계는 다음과 같다. 먼저, 검색식을 통해 DB로부터 수집한 특허 문서 데이터에서 명칭, 요약, 청구항을 포함한 텍스트 정보를 추출한다. 다음으로 Word2Vec과 Textmining 기법 활용하여 거리 행렬 및 문서-단어 행렬(Document-Term Matrix, DTM)을 구축한다. 두 행렬을 사용하여 가중치 행렬을 생성하고 검색식 키워드와 비교하여 노이즈 특허 Seed label을 생성한다. 마지막으로, 이를 Random forest로 학습 및 분류한다.

3.2 가중치 행렬 구축

본 연구에서 제안하는 모형은 Word2Vec을 활용하여 가중치 행렬 구축 후, 노이즈 특허 Seed label을 생성한다. <그림 2>는 해당 과정을 도식화한 것이다.

추출된 텍스트 정보에 Word2Vec을 활용하여 벡터 공간에 Embedding 한다. 다음으로, 단어 간의 Euclidean distance를 측정하여 거리 행렬을 구축한다.



<그림 2> 가중치 행렬 구축 과정

앞서 추출한 텍스트 정보에 Textmining 기법을 적용하여 전처리를 수행하고 DTM을 구축한다. 또한, DTM과 앞서 구축한 거리 행렬을 행렬곱(Matrix multiplication)하여 가중치 행렬을 구축한다.

검색식에 사용된 단어는 주로 분석 목적과 일치하는 기술 키워드로 구성되어 있다. 따라서 검색식 키워드가 많이 등장한 것은 유효특허, 상대적으로 적은 것은 노이즈 특허일 가능성이 높다. 이러한 이유로, 제안하는 연구 모형은 검색식에 사용된 키워드와 일치하는 단어를 이용하여 노이즈 특허 Seed label을 생성한다.

가중치 행렬에서 검색식 키워드와 일치하는 단어

열(Words Columns)만을 추출하고 각 열의 값을 0~1로 정규화(Normalization)한다. 다음으로, 정규화된 값의 합을 통해 문서별 점수를 도출한다. 하위 점수를 보유한 20건은 노이즈 특허로, 상위 점수를 보유한 20건은 유효 특허로 Seed label을 부여한다.

3.3 노이즈 특허 분류

본 연구에서는 Seed label을 부여한 특허들에서 더욱 많은 정보를 추출하기 위해 앞서 구축한 DTM을 활용한다. DTM에 Seed label을 부여한 총 40건(노이즈 특허 20건, 유효특허 20건)을 Training data로써, Random forest로 학습한다. 나머지 데이터에 학습된 모형을 사용하여 노이즈 특허분류를 수행한다. 또한, 분류 성능 비교를 위하여 다양한 알고리즘으로 분류한 결과를 함께 제시한다.

Ⅳ. 실험 및 결과

4.1 실험 데이터 수집

본 연구에서 제안하는 모형의 분류 성능 확인을 위해 다음과 같은 특허 데이터를 수집하여 실험을 진행한다. 수집 대상 데이터는 해양감시 네트워크 (Ocean Surveillance & Tracking Network, OSTN) 기술 관련 특허 중 미국 특허청(United States Patent and Trademark Office, USPTO)에 공개 및 등록된 것이다. 검색기간은 1992년 1월부터 2014년 2월로 제한하며, 특허 DB는 윕스를 사용한다. 모형의 분류 성능 측정을 위한 노이즈 특허 label은 e특허나라의 OSTN 동향보고서에서 관련 전문가들이 직접 분류한 특허 목록을 활용한다.

검색식을 구성하여 수집한 특허는 총 580건이다. 이 중 노이즈 특허는 464건(80%), 유효특허는 116건(20%)이다. <표 1>은 노이즈 특허 Seed label 생성을 위해 사용한 검색식 키워드 목록을 나타낸다. 본 실험에서는 <표 1>의 키워드와 일치하는 가중치 행렬의 단어열(Words Columns)만을 추출하여 노이즈 특허 Seed label을 생성한다.

<표 1> 검색식 키워드 목록

No.	Keyword	No.	Keyword
1	acoust	11	moor
2	communic	12	network
3	demodul	13	pollut
4	extern	14	protocol
5	hybrid	15	rout
6	identif	16	secur
7	local	17	sensor
8	locat	18	submarin
9	modul	19	transceiv
10	monitor	20	vehicl

4.2 노이즈 특허 분류 결과

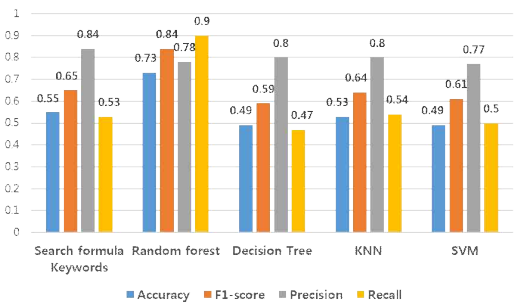
수집된 데이터에서 Textmining 기법을 사용하여 단어를 추출한 결과 총 3758개였다. 이를 Word2Vec 방법 중 Skip-gram을 사용하여 100차원의 공간에 Embedding 하였다.

거리 행렬과 DTM을 행렬곱하여 도출한 가중치 행렬의 크기는 580X3758이며, <표 1>의 검색식 단어와 일치하는 단어열만을 추출하였다. 다음으로 해당 열의 값들을 정규화하고, 문서별로 단어 값의 합을 구하여 총 40건의 Seed label을 부여하였다. 아래의 <표 2>는 Seed label의 특허목록을 나타낸다.

<표 2>에서 왼쪽 열은 노이즈 특허로, 오른쪽 열은 유효특허로 Seed label이 부여된 것이다. 상기의 특허들을 Training 데이터로써, 학습하여 분류한 결과는 <그림 3>과 같다.

<표 2> Seed label 특허목록

Application number	Seed label	Application number	Seed label
13/417501	Noise	09/199832	Validity
13/417546		09/063728	
13/417526		12/802018	
11/959537		09/341115	
13/316719		10/902958	
13/236321		12/642065	
11/195250		08/488840	
29/122019		09/569696	
07/953169		11/183308	
12/373010		07/878697	
12/543306		08/723534	
08/965262		09/417977	
29/241679		10/358577	
08/915393		08/797976	
13/417552		11/602429	
09/882329		08/181185	
13/417563		11/709404	
08/789065		11/795835	
29/241680		08/842602	
13/417534		08/400600	



<그림 3> 노이즈 특허 분류 결과

<그림 3>에서 Accuracy는 실제 전문가들이 분류한 label과의 일치 정도를 나타낸 것이며, F1-score는 정밀도(Precision)와 재현율(Recall)의 조화평균이다. Accuracy와 F1-score 모두 값이 높을수록 분류 성능이 우수한 것을 의미한다.

제시한 결과는 분류 알고리즘별로 Accuracy와 F1-score를 모두 나타내었다. 그중 Search formula Keywords는 Seed label를 생성할 시에 사용한 문서별 점수로만 분류한 것이다. 분류 결과, Random forest를 적용한 것이 73%의 Accuracy를 보인 것으로 가장 우수하였다. 또한, 다른 방법들은 대부분 비슷한 분류 성능을 보였다.

Ⅴ. 결론 및 향후 연구

본 논문에서는 효율적인 특허정보 조사를 위한 노이즈 특허 분류 모형을 제안하였다. 해당 모형은 Word2Vec과 가중치 행렬을 활용하여 Seed label을 생성하므로, 학습에 필요한 분류된 label이 요구되지 않는다. 또한, 특허 데이터의 정보손실 방지를 위해 Seed label의 전체 단어를 Bootstrap 기반의 Random forest로 학습 및 분류한다.

실험 데이터로는 OSTN 기술 관련 특허 중 USPTO에 공개 및 등록된 것을 사용하였으며, 특허 검색 DB는 윕스를 이용하였다. 분류 성능 측정을 위한 노이즈 특허 label은 e특허나라의 OSTN 동향보고서에서 관련 기술 전문가들이 직접 분류한 특허목록을 사용하였다.

수집한 특허는 총 580건이었으며, 추출된 단어는 3758개였다. Word2Vec을 활용하여 생성한 거리 행렬과 DTM을 행렬곱한 가중치 행렬의 크기는 580X3758이었다. 이를 검색식 키워드 20개와 매칭되는 단어열만을 추출하여 총 40개의 Seed label을 생성하고 Random forest로 학습 및 분류하였다.

분류 성능 비교를 위해 Decision Tree, KNN, SVM을 포함하는 다양한 알고리즘의 결과를 함께 제시하였다. 대부분 분류 알고리즘들의 성능이 유사하였으며, 제안한 모형이 실제 전문가들이 분류한 label과 73%의 일치율을 보인 것으로 가장 우수하였다.

본 연구에서는 노이즈 특허 Seed label 생성을 위한 거리 행렬을 Euclidean distance만을 사용하여 구축하였다. 이는 노이즈 특허분류를 위해 단어의 가중치를 부여하는 작업이다. 따라서, 향후 연구에서는 다

양한 거리 측도를 활용하여 가장 적합한 가중치 행렬 생성 방법에 관한 연구가 필요할 것으로 보인다.

참고문헌

[1] D. Hunt, L. D. Nguyen, M. Rodgers, Patent Searching Tools & Techniques, Wiley, New Jersey, 2007.

[2] A. Abbas, L. Zhang, S. U. Khan, "A literature review on the state-of-the-art in patent analysis," World Patent Information, Vol. 37, 2014, pp.3-13.

[3] S. Jun, "A Big Data Learning for Patent Analysis," Journal of Korean Institute of Intelligent Systems, Vol. 23, No. 5, 2013, pp.406-411.

[4] Korean Intellectual Property Office (KIPO), Korean Invention Promotion Association (KIPA), Patent and Information Analysis (for Researchers), KIPO, Seoul, 2006.

[5] Korean Intellectual Property Office (KIPO), Korean Intellectual Property Strategy Agency (KISTA), Intellectual Property Research & Development, KIPO, Seoul, 2012.

[6] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.

[7] X. Rong, "word2vec Parameter Learning Explained," arXiv preprint arXiv:1411.2738, 2016.

[8] S. Goki, Deep Learning from Scratch 2, O'Reilly, California, 2019.

[9] W. Kim, D. Kim, H. Jang, "Semantic Extention Search for Documents Using the Word2vec," The Journal of the Korea Contents Association, Vol. 16, No. 10, 2016, pp.687-692.

[10] D. Seo, K. H. Mo, J. Park, G. Lee, P. Kang, "Word Sentiment Score Evaluation based on Graph-Based Semi-Supervised Learning and Word Embedding," Journal of the Korean Institute of Industrial Engineers, Vol. 43, No. 5, 2017, pp.330-340.

[11] K. Kim, C. Park, "Automatic IPC Classification of Patent Documents Using Word2Vec and Two Layers Bidirectional Long Short Term Memory Network," The Journal of KINGComputing, Vol. 15, No. 2, 2019, pp.50-60.

[12] J. Jeong, K. H. Mo, S. Seo, C. Y. Kim, H. Kim, P. Kang, "Unsupervised Document Multi-Category Weight Extraction based on Word Embedding and Word Network Analysis: A Case Study on Mobile Phone Reviews," Journal of the Korean Institute of Industrial Engineers, Vol. 44, No. 6, 2018, pp.442-451.

[13] A. Geron, Hands-on Machine Learning with Scikit-learn & Tensorflow, O'Reilly, California, 2017.

[14] C. Park, Y. Kim, J. Kim, J. Song, H. Choi, R Data Mining, Kyowoosa, Seoul, 2011.

[15] V. Sventnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. "Feuston, Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," Journal of Chemical Information and Computer Science, Vol. 43, No. 6, 2003, pp.1947-1958.

[16] W. T. Aung, Y. Myanmar, K. H. S. Hla, "Random forest classifier for multi-category classification of web pages," Proceedings of 2009 IEEE Asia-Pacific Services Computing Conference (APSCC), 2009, pp.372-376.

[17] A. Onan, S. Korukoglu, H. Bulut, "Ensemble of

keyword extraction methods and classifier in text classification," Expert Systems with Applications, Vol. 57, No. 15, 2016, pp.232-247.

■ 저자소개 ■

김 영 호
Kim, Youngho

2016년 9월~현재
고려대학교 산업경영공학부
석·박사통합과정
2016년 8월 목원대학교 경영정보학과
(경영정보학사)

관심분야 : Patent Analysis, Text-mining, Machine learning, etc.
E-mail : youngho0928@korea.ac.kr

박 상 성
Park, Sangsung

2018년~현재
청주대학교 소프트웨어융합학부
조교수
2015년~2018년
고려대학교 기술경영전문대학원
조교수
2006년~2014년
고려대학교 산업경영공학부 연구교수
2006년 고려대학교 산업시스템정보공학과
(공학박사)
관심분야 : Patent Analysis, Data Mining, Management of Technology, Technology Evaluation
E-mail : hanyul@cju.ac.kr

장 동 식
Jang, Dongsik

1989년~현재
고려대학교 산업경영공학부 교수
1988년 텍사스A&M 산업공학과 (공학박사)
1985년 텍사스주립대학교 산업공학과
(공학석사)
1979년 고려대학교 산업공학과 (공학사)

관심분야 : Project Management, Pattern Recognition, Data Mining
E-mail : jang@korea.ac.kr

논문접수일 : 2019년 11월 18일
게재확정일 : 2019년 12월 4일