JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 22, No. 8: 1049~1054, Aug. 2018

잠재 의미 분석을 적용한 유사 특허 검색 서비스 시스템

임현근[1] · 김재윤[1] · 정회경[1*]

Similar Patent Search Service System using Latent Dirichlet Allocation

HyunKeun Lim[1] · Jaeyoon Kim[1] · Hoekyung Jung[1*]

[1*]Department of Computer Engineering, Paichai University, Daejeon 35345, Korea

요 약

유사 특허를 검색하는 방법으로 기존에는 키워드 검색 방법을 사용하고 최근에는 머신러닝을 활용한 자동분류 방법을 사용하고 있다. 키워드 검색은 데이터 정제를 통해 정형화된 데이터 분석 방법으로 단문일 경우 검색에서는 정확도는 높지만 문서와 같이 여러 단어로 이루어진 장문일 경우 문장에 내포된 의미 분석을 할 수 없었다. 의미 분석 단계에서의 자동 분류 방법은 비정형 데이터 분석 방법으로 여러 단어로 이루어진 문장을 분류하는데 사용되고 있다. 그 동안 두 가지 방법을 결합하여 유사 문서 검색을 하려는 시도가 있었지만 비정형 데이터와 정형 데이터의 동시 사용에는 분석하는 방법이 다르기 때문에 동시 적용에는 알고리즘 상의 문제가 있었다. 이에 본 논문에서는 문서에서 함축된 키워드를 검출하고 잠재 의미 분석(LDA) 방식을 사용하여 사람이 개입하지 않고 문서를 효율적으로 자동분류하고 유사 특허를 검색할 수 있는 방법을 연구하였다.

ABSTRACT

Keyword searching used in the past as a method of finding similar patents, and automated classification by machine learning is using in recently. Keyword searching is a method of analyzing data that is formalized through data refinement. While the accuracy for short text is high, long one consisted of several words like as document that is not able to analyze the meaning contained in sentences. In semantic analysis level, the method of automatic classification is used to classify sentences composed of several words by unstructured data analysis. There was an attempt to find similar documents by combining the two methods. However, it have a problem in the algorithm w the methods of analysis are different ways to use simultaneous unstructured data and regular data. In this paper, we study the method of extracting keywords implied in the document and using the LDA(Latent Semantic Analysis) method to classify documents efficiently without human intervention and finding similar patents.

키워드 : 머신러닝, 문서분류, 유사특허검색, 잠재의미분석, 키워드추출

Keyword : Machine Learning, Document Classification, Similar Patent Search, LDA, Keyword Extract

Received 9 May 2018, **Revised** 23 May 2018, **Accepted** 4 June 2018

*** Corresponding Author** Hoekyung Jung(E-mail:hkjung@pcu.ac.kr, Tel:+82-42-520-5640)
Department of Computer Engineering, Paichai University, Daejeon 35345, Korea

Open Access http://doi.org/10.6109/jkiice.2018.22.8.1049 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

본 논문에서는 특허문서를 기계 학습에 의해 국제 특허 분류(IPC) 기준에 맞게 자동으로 분류하고 유사한 특허를 검색하는 시스템에 관한 것으로 베이지안(Baysian) 확률론을 이용한 기술 주제 분류와 잠재 의미 분석 기법을 이용하여 유사 특허 검색의 효율성을 높일 수 있는 방법을 제안한다. 종래 기계 학습을 이용한 유사 특허 문서 검색은 용어 빈도를 고려한 자질 선택 기법에 의존해서 특허를 분류하고 메타 키워드의 가중치를 이용한 검색 방법이었으나, 현대 기술 용어의 발생 속도와 표현의 다양성을 고려할 때 특허 문서들 간의 관련성을 분석하는데 있어서는 용어 자체의 빈도 보다는 용어의 개념에 의한 접근이 보다 효과적일 것이라 판단하였다.

이에 종래 방법과 잠재된 의미 유사도를 비교하는 방법을 결합하여 사용하는 방법을 제안하였다. 기존의 분류 알고리즘인 Naive 베이지안 알고리즘의 자질 선택 기법으로 카이제곱 테스트(Chi-Square Test)를 이용하였고 문서에서 키워드 추출은 Text Rank 기법을 사용하였다. 마지막으로 잠재 의미 추론 기법으로 LDA 알고리즘을 사용하여 특허 의미 유사도를 측정하였다. 서비스 범위는 국제, 미국, 유럽, 일본, 중국으로 확대하기 위해서 모든 문서는 영어로 번역해서 사용한다.

분석 언어를 영어로 사용하는 또 다른 이유로는 영어는 어절 단위의 의미를 가지는 언어이기 때문에 형태소 분석에서 키워드를 추출하는데 유리하며 Wordnet을 사용할 수 있어서 추후 다양한 유사 단어의 자질 추출이 가능하기 때문이다.

Ⅱ. 관련 연구

2.1. 특허 IPC 구조

국제특허분류(IPC, International Patent Classification)는 특허 문서들을 고유의 분류코드로 할당한다. 이 분류코드는 특허문서에 대하여 국제적으로 통일되게 분류하기 위한 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 그림 1과 같은 계층적 구조를 가진다

예를 들어 ‘G02B 6/44’의 IPC코드는 G 섹션, G02클래스(광학), G02B 서브클래스(광학요소), G02B 6/00 메인그룹(광학파관), G02B 6/44 서브그룹(광섬유케이블)에 해당한다. 이와 같이 IPC레벨이 아래로 내려갈수록 그 기술 분야는 더욱 세분화된다.

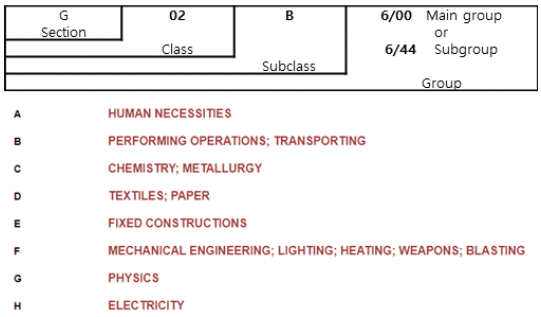


Fig. 1 IPC Code Structure

2.2. 특허 문서 분류 방법

제한된 학습데이터를 가지고 IPC 분류과정의 효율성을 높이기 위하여 지도학습 방법 중 Naive 베이지안 기계학습 알고리즘 기반의 IPC 자동 분류를 사용한다. 데이터 소스는 특허문서의 초록 필드를 사용하여 동시에 다수의 IPC 분류코드를 가지는 다중 레이블 분류(multi-label classification) 모델을 구축한다.

분류 알고리즘을 적용하기 위한 전처리 과정으로 특징을 추출하는 과정에서 다양한 방법들이 고안되었고, 여기서는 자질 선택 방법으로 기존 사용하던 N-gram 방식보다 개선된 방법으로 카이제곱 분포 통계 기법을 사용하였다[1].

2.3. 키워드 추출

분석 대상 특허를 검색하기 위한 용도로 사용할 키워드 추출 방법으로 Text Rank 알고리즘을 사용하여 분석 대상이 되는 문서에서 중요도가 높은 키워드를 추출한다[2]. TextRank는 텍스트에 관한 graph-based ranking model로, Google의 PageRank를 활용한 알고리즘이다[3].

PageRank 알고리즘은 하이퍼링크를 가지는 웹 문서에 상대적 중요도에 따라 가중치를 부여하는 방법으로 서로간의 인용과 참조로 연결된 임의의 묶음에 적용할 수 있다. PageRank가 높은 웹페이지는 다른 웹 사이트

로 부터 링크를 받는다. 즉 다른 사이트가 참조를 많이 한 것으로 해석할 수 있다. TextRank 방식으로 추출된 키워드에 상호 동시 출연 빈도수 Pointwise Mutual information(PMI) 값이 높은 키워드를 추출한다.

2.4. LDA 유사도 검증

문서의 잠재 의미 분석을 위한 방법으로 토픽 모델링(Topic Modeling) 기법 중 LDA 알고리즘을 사용하였다. 기계 학습 및 자연언어 처리 분야에서 토픽 모델이란 문서 집합의 추상적인 "주제"를 발견하기 위한 통계적 모델 중 하나로, 텍스트 본문의 숨겨진 의미구조를 발견하기 위해 사용되는 텍스트 마이닝 기법 중 하나이다. 현재 사용되는 가장 일반적인 주제 모델링 방법인 LDA는 여러 주제가 혼합된 문서를 다룰 수 있게 한다. LDA의 아키텍처, 즉 LDA가 가정하는 문서생성과정은 그림 2와 같다. D는 말뭉치 전체 문서 개수, K는 전체 토픽 수(하이퍼 파라메터), N은 d번째 문서의 단어 수를 의미한다.

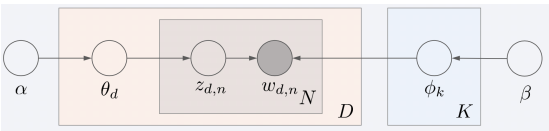


Fig. 2 LDA Algorithm

관찰 가능한 변수는 d번째 문서에 등장한 n번째 단어 W(d,n)이 유일하다. 임의로 지정된 하이퍼 파라메터 α, β를 제외한 모든 잠재 변수를 추정해야 한다. 확률과정과 결합 확률을 각각 그림 2와 다음 수식으로 나타내었다. 여기서 Gibbs Sampling 과정을 통해 p(z,ϕ,Θ|w)를 최대로 만드는 z,ϕ,Θ를 찾는다.

$$p(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\phi_i|\beta) \prod_{d=1}^{D} p(\theta_d|\alpha) \left\{ \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\phi_{1:K}, z_{d,n}) \right\}$$

표 1은 임의 6개의 문서에 대한 Topic 3개의 Θ값이다. 문서에 대한 Θ 값의 합은 분포 확률이기 때문에 항상 1이 된다. 문서의 유사도를 측정하기 위한 측정 방법은 비교 대상 문서와 다른 문서들의 Θ 값을 Cosine 유사도로 비교한다.

Table. 1 LDA Theta Distribution

Docs	Topic 1	Topic 2	Topic 3
Doc 1	0.400	0.000	0.600
Doc 2	0.000	0.600	0.400
Doc 3	0.375	0.625	0.000
Doc 4	0.000	0.375	0.625
Doc 5	0.500	0.000	0.500
Doc 6	0.500	0.500	0.000

Ⅲ. 시스템 구성

3.1. 서비스 구성

시스템 구성은 Springboot 기반의 웹 어플리케이션 서비스로 개발한다. 이는 그림 3과 같다. 내부 모듈 구성으로 기본적인 화면 입출력 UI 서비스 이외에 내부적으로 Naive bayesian classification 모듈, Text Rank 키워드 추출 모듈, LDA 잠재의미분석 모듈 등 3가지로 구성하며 외부 기능으로 키워드 파싱, 번역, 언어 탐지 API 기능을 사용한다. 또한 특허 검색 API로 NOS Open API 모듈을 사용한다.

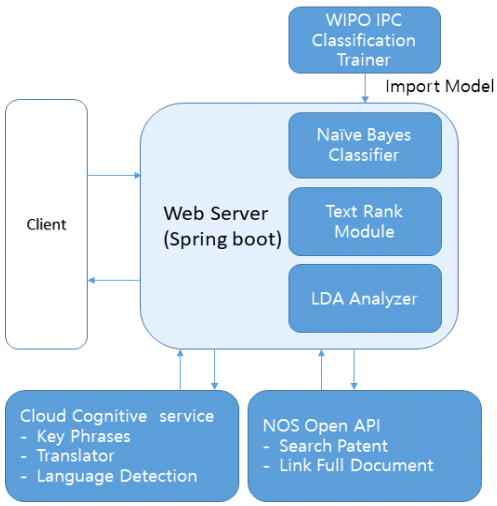


Fig. 3 System Structure

3.2. 단계별 처리 과정

전체 4단계의 과정으로 문서 전처리, 문서 분석, 특허 검색, 유사 특허 검색 4단계의 과정으로 구성된다. 이는 그림 4와 같다.

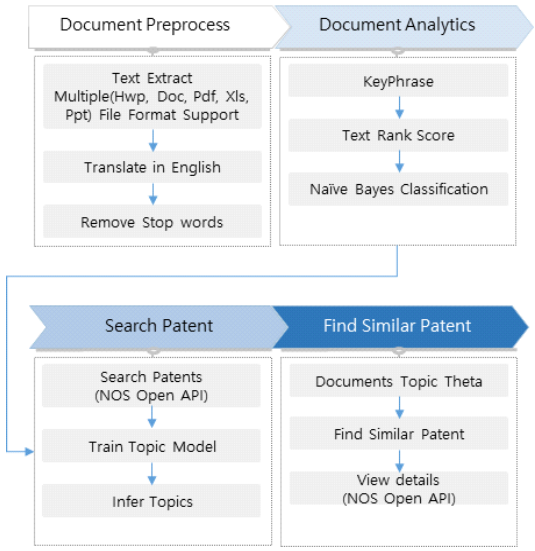


Fig. 4 System Working Process

1단계) 문서의 전처리

전처리 첫 번째 과정으로 문서를 입력받아 텍스트를 추출한다. 분석 대상은 텍스트로만 한정하기 때문에 문서에서 텍스트만 추출한다. 다음 단계로 분석 대상의 언어를 구분하고 영어로 번역하는 과정이다.

-Text Translate

> 풍력 발전은 풍력 터빈을 이용해서 바람(풍력)을 전력으로 바꾸는 일이다.
>
> Wind power is a wind turbine and wind (wind) using the power is changing.

언어 확인과 번역은 Google 번역 API를 활용하였다. 언어 확인 API 요청을 할 경우 다음과 같이 0과 1사이의 값을 반환한다. 마지막으로 형태소 분석을 통해 필요한 단어만 추출하는 과정으로 Stop word 필터를 적용하여 불필요한 단어들을 제외하도록 하였다.

2단계) 문서 분석 단계

두 가지 알고리즘을 처리하는 과정이다. 하나는 Text Rank 알고리즘을 통해 키워드를 추출하는 과정이고 다른 하나는 Naive Bayesian을 이용한 특허 분류를 하는 과정이다. 키워드 추출은 앞 장에서 설명했듯이 분석 대상 텍스트에서 동시 출현 빈도가 가장 높은 어절을 추출한다. 학습용 분류 데이터를 이용하여 분류 모델을 구성한다. WIPO에서 제공하는 75000개의 영어 기준 학습 데이터를 사용하여 모델을 구성하였다. 이는 그림 5와 같다.

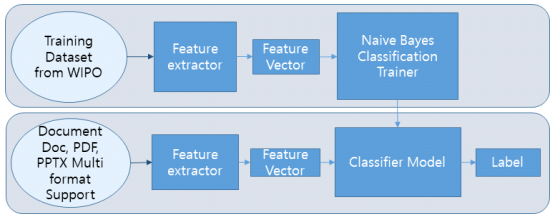


Fig. 5 Naive Bayesian Document Training and Running

학습된 분류 모델을 통해 분석 대상을 적용하여 분류한다. 분류된 결과는 3레벨 128개의 클래스로 분류된다. 결과 데이터는 상대값 비교로 Jaccard 거리값으로 계산하며 결과는 최대 1, 최소 0으로 정규화 한 결과로 변환한다.

-Text Source

> Wind power is a wind turbine and wind (wind) using the power is changing. Today, wind power is relatively cheap in many countries, renewable energy provides electricity to carbon is almost a circle.

-Result of IPC Classification

분	분류...	설명
1	1.0	액체용기계 또는 기관 (액체 및 압축성유체용 F01 ; 액체용용적...
2	0.995	전력의 발전, 변환, 배전
3	0.704	기본적 전기소자
4	0.655	기계 또는 기관일반(연소기관 F02 ; 액체용 F03, F04 ; ;기관설...
5	0.651	연소기관(그것을 위해서 주기적으로 작동하는 밸브, 윤활, 배기...
6	0.536	제어 ; 조정 (특수한 용도에 특히 적합한 것은, 그 분야의 관련 ...

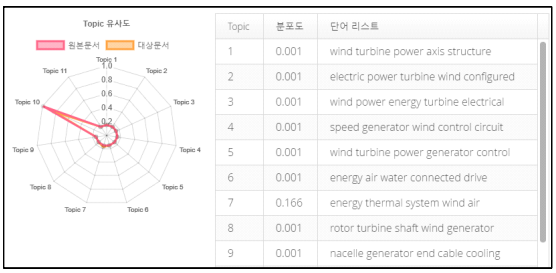
3단계) 특허 검색

NDSL Open Service(NOS)에서 제공하는 Open API를 사용하였다. 이는 국가과학기술정보센터(NDSL)에서 학술용으로 제공하는 기능으로 국내/국외 특허 검색 서비스를 제공하고 있다. Open API를 사용하여 효과적으로 데이터를 가져오기 위해 검색 쿼리 작성기를 별도로 구현하였다.

4단계) 특허 토픽 분석

LDA 알고리즘을 사용하여 특허 문서들의 Topic을 분석한다. 분석된 결과에 따라 분석 대상과의 Topic 유사도 결과값을 도출한다. 인자값으로 $\alpha = 0.01$,$\beta = 0.01$,

문서생성회수=500, 샘플링 횟수= 500, Topic = 11개를 사용한다. 수치 테스트를 위한 임의의 기준값이며 추후 최적값에 대한 연구는 별도로 진행한다. 분석 대상과 특허 문서들의 Θ값으로 유사도를 측정하고 결과는 Chart.js을 사용하여 스타 모양의 Radar chart로 출력한다. 이는 그림 6과 같다.



Topic	분포도	단어 리스트
1	0.001	wind turbine power axis structure
2	0.001	electric power turbine wind configured
3	0.001	wind power energy turbine electrical
4	0.001	speed generator wind control circuit
5	0.001	wind turbine power generator control
6	0.001	energy air water connected drive
7	0.166	energy thermal system wind air
8	0.001	rotor turbine shaft wind generator
9	0.001	nacelle generator end cable cooling

Fig. 6 LDA Topic Similarity

Ⅳ. 실험

유사 문서 검색에 대한 정확한 분류 측정의 객관적인 기준을 위하여 실험 과정은 두 단계로 구분해서 진행한다. 첫 번째로는 IPC분류 정확도를 측정하고 두 번째로는 LDA 문서 유사도 검증 테스트를 진행하도록 한다. Naive 베이지안 분류 정확도 테스트를 위하여 WIPO에서 제공하는 분류 학습용 데이터 이외에 테스트 셋을 사용하였다. 사용한 테스트 모듈은 Naive 베이지안 알고리즘을 사용하였다.

분류 학습용 인자값으로 카이스퀘어(CHI) 값은 0.05를 사용하고 영문을 제외한 숫자와 기호, 외국어는 제외했다. 불용어(Stopword)는 전치사 대명사 등을 포함 총 420개를 사용하였다. 75000개의 학습용 특허 데이터를 사용하였고 i5-1.8Gz 2 core에서 학습 시간은 10분 소요됐다. 테스트 데이터는 A그룹 5160개의 특허를 사용했다. 총 분석 시간은 3분 걸렸다. 측정 방법은 테스트 특허의 분류 결과값 0-1 사이 값의 전체 평균으로 계산하였다. 실험에서 평균 특허 분류 정확도는 88.54%로 나왔다.

K-mean Naive 베이지안을 사용한 Single Match 정확도는 87.2%로 나왔다. Multi Match 정확도는 이보다 5% 낮다[4]. CHI를 사용한 SVM 알고리즘의 측정치 최고 수치가 88.7% 나왔다[5]. 위의 실험 결과 CHI를 사용한 Naive 베이지안은 SVM와 동일한 결과를 보인다.

두 번째로 LDA 문서 유사도 검출 테스트를 위하여 테스트 데이터는 임의로 50개의 특허 데이터의 이름, 요약, 청구항을 사용 하였다. 문서 잠재 유사도 측정 방법으로 정규 상호정보(NMI, Normalized Mutual Information)를 측정하였다. 이는 Clustering이 얼마나 잘되었는지에 대한 평가 지표로 아래의 식에 따른다[6].

$$NMI(X,Y) = \frac{2I(X;Y)}{H(X)+H(Y)}$$

I(X;Y)는 X와 Y의 상호정보(Mutual information)값이다. X는 X={X1, X2...Xn}, Y는 Y={Y1, Y2...Yn}으로 Xi 는 LDA Topic i에 해당하는 데이터 집합이다. Yj는 j번째 라벨을 가지고 있는 데이터이다. 결국 Label j에 해당하는 Topic j의 분포 확률이 얼마나 되는지 측정하는 것이다. NMI = 0 이면 클러스터링 결과가 문서와 완전히 다르다는 의미이고 NMI = 1은 문서들의 고유 특징을 가진다는 의미이다.

위의 샘플 데이터와 50개의 데이터셋을 가지고 테스트시 평균 NMI = 0.7468의 값을 얻었다. 분석 대상에 대한 Purity Accuracy = 0.2156 이다. K-mean 평균 문서 클러스터링 결과 9개의 군집 성공률 77% 비교 시 Topic 의미 분석에 대한 결과로 의미가 있다고 본다[7].

Ⅴ. 결론

Naive bayesian 알고리즘으로 SVM하고 동일한 결과를 얻을 수 있다는 것을 확인하였고, LDA 알고리즘을 사용하여 다중 의미 분석까지 하였다.

A4 한장 분량의 문서를 분석하는데 Intel i5 3세대 노트북에서 20초 정도 소요되었다. 사용 메모리도 웹 서비스에서 4G 정도 사용하였다. 더 정확한 값을 도출하기 위해 특허 분류를 한 단계 더 높이면 메모리는 10G 정도 필요로 하다. 상대적으로 유사 특허 검색의 다른 방식으로 Google 특허 서비스에서는 특허 전체를 Doc2Vec을 사용하고 있다. 미국 특허로만 구성한 벡터 모델의 용량이 10G 정도이고 시스템에서 실제 서비스로 사용하기에는 고 사양 컴퓨터가 요구된다.

본 논문에서 사용한 두 가지 알고리즘을 활용하여 저사양 컴퓨터에서 효과적으로 유사 특허 검색 시스템을

구현할 수 있었다. 또한 기존 LDA를 보완한 LDA K-mean 알고리즘 사용으로 기존 NMI = 0.74에서 평균 NMI = 0.85 까지 높이고 AU-ROC 정확도 테스트에서 60% 에서 73% 까지 높일 수 있었다.

향후연구로는 두 가지 알고리즘을 하나로 만드는 것으로 LDA 방식에 Naive bayesian의 특징을 Topic으로 사용할 수 있는가에 대한 추가 연구가 필요하다.

REFERENCES

[1] Suhendra, I. Ranggadara, “Naive Bayes Algorithm with Chi Square and NGram Feature for Reviewing Laptop Product on Amazon Site,” *International Research Journal of Computer Science*, Issue 12, vol. 4, pp. 28-33, Dec. 2017.

[2] J. W. Lee, I. S. Kang, H. K. Jung, “XML Document Keyword Weight Analysis based Paragraph Extraction Model,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 21, no. 11, pp. 2133-2138, Nov. 2017.

[3] K. H. Song, Y. S. Kim, “Automatic Keyword Extraction using Hierarchical Graph Model Based on Word Co-occurrences,” *Journal of Korean Institute of Information Scientists and Engineers*, vol. 44, no. 5, pp. 522-536, May. 2017.

[4] S. R. Lim, Y. J. Kwon, “IPC Multi-label Classification based on Functional Characteristics of Fields in Patent Documents,” *Journal of Internet Computing and Services*, vol. 18, no. 1, pp. 77-88, Feb. 2017.

[5] T. H. Jeen, “Patent documents automatic classification with dimension reduced features using latent semantic analysis,” M. S. dissertation, Computer and Information Technology, Korea University, Feb. 2014.

[6] R. Mehrotra, S. Sanner, W. Buntine, L. Xie, “Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling,” *ACM Special Interest Group on Information Retrieval*, pp. 889-892, Jul. 2013.

[7] W. S. Kim, S. Y. Kim, “Document Clustering Technique by K-means Algorithm and PCA,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 18, no. 3, pp. 625-630, Mar. 2014.

임현근(Hyunkeun Lim)

2002년 세명대학교 컴퓨터과학과(공학사)
2018년 배재대학교 컴퓨터공학과 석사과정(공학석사)
2003년 ~ 2005 Efusion 독일 모바일게임 개발자
2005년 ~ 2010년 Mbizglobal 모바일 솔루션 수석개발자
2015년 ~ 2018년 ㈜유진정보 데이터분석팀 차장
※관심분야 : 텍스트마이닝, 모바일게임, IoT

김재윤(Jaeyoon Kim)

2017년 배재대학교 컴퓨터공학과(공학석사)
2018년 ~ 현재 배재대학교 컴퓨터공학과 박사과정
※관심분야 : BigData, Block Chain, IoT, 텍스트마이닝

정회경(Hoekyung Jung)

1985년 광운대학교 컴퓨터공학과(공학사)
1987년 광운대학교 컴퓨터공학과(공학석사)
1993년 광운대학교 컴퓨터공학과(공학박사)
1994년 ~ 현재 배재대학교 컴퓨터공학과 교수
※관심분야 : 멀티미디어 문서정보처리, Ubiquitous Computing, USN, IoT, BigData, Embedded System