

<https://doi.org/10.7236/JIIBC.2022.22.5.171>
JIIBC 2022-5-25

학습률 향상을 위한 딥러닝 기반 맞춤형 문제 추천 알고리즘

Deep learning-based custom problem recommendation algorithm to improve learning rate

임민아*, 황승연**, 김정준***

Min-Ah Lim*, Seung-Yeon Hwang**, Jeong-Jun Kim***

요약 최근 딥러닝 기술의 발전과 함께 추천 시스템의 영역도 다양해졌다. 본 논문은 학습률 향상을 위한 알고리즘을 연구하였으며 Word2Vec 모델의 성능 특징과 비교를 통해 단어에 따른 유의어 결과를 연구하였다. 문제 추천 알고리즘은 Word2Vec 모델의 특징인 텍스트 간 의미 반영 및 유사성 테스트를 통해 표현된 값으로 구현됐다. Word2Vec의 학습 결과를 통해 텍스트 유사도 값을 이용해 문제 추천을 진행하였으며 유사도가 높은 문제를 추천할 수 있다. 실험 과정에서 정량적인 데이터양으로는 정확성이 낮아지는 결과를 보았으며 데이터 셋의 데이터양이 방대할수록 정확성을 높일 수 있음을 확인하였다.

Abstract With the recent development of deep learning technology, the areas of recommendation systems have also diversified. This paper studied algorithms to improve the learning rate and studied the significance results according to words through comparison with the performance characteristics of the Word2Vec model. The problem recommendation algorithm was implemented with the values expressed through the reflection of meaning and similarity test between texts, which are characteristics of the Word2Vec model. Through Word2Vec's learning results, problem recommendations were conducted using text similarity values, and problems with high similarity can be recommended. In the experimental process, it was seen that the accuracy decreased with the quantitative amount of data, and it was confirmed that the larger the amount of data in the data set, the higher the accuracy.

Key Words : Deep learning, Recommender system, Text similarity, Word2Vec

1. 서론

인공지능의 발전은 현재 큰 관심을 받는 분야이며 추천 시스템은 딥러닝 기술을 기반으로 누구나 일상생활에

서 사용할 수 있게 상용화가 진행되고 있다.

과거에는 온라인보다 오프라인으로 직접 상품들을 비교해가며 구매하였다. 즉 판매자에게 직접 추천받거나 지인들의 추천을 받아가면 구매한 것이다. 그러나 최근

*준회원, 안양대학교 소프트웨어학과

**준회원, 안양대학교 컴퓨터공학과

***정회원, 안양대학교 소프트웨어학과

접수일자 2022년 7월 18일, 수정완료 2022년 9월 8일
게재확정일자 2022년 10월 7일

Received: 18 July, 2022 / Revised: 8 September, 2022 /

Accepted: 7 October, 2022

***Corresponding Author: jkim@anyang.ac.kr

Dept. ICT Convergence Engineering at Anyang University, Korea

에 온라인 쇼핑몰의 추천 시스템 알고리즘으로 비슷한 상품들을 추천받아 소비자는 복잡한 과정 없이 편리하게 상품을 구매할 수 있도록 유도한다. 이렇게 구매한 상품에 대해서 구매자는 만족도를 작성할 수 있으며 이를 고려해 다른 소비자들에게 상품을 추천할 수 있다. 이러한 추천 시스템은 사용자에게 방대한 정보의 혼란성을 줄여 주고 효율성을 제공한다.

추천 시스템은 크게 협업적 필터링 기반 알고리즘과 콘텐츠 기반 알고리즘으로 나뉜다. 먼저 협업적 필터링은 앞서 말한 상품 판매와 구매에서 고객의 만족도 또는 영화 만족도 등 정보 기반으로 추천하는 방식이다. 콘텐츠 기반 알고리즘은 본 연구에서 사용한 알고리즘으로 벡터 기반으로 메타데이터를 분석하여 학습을 위한 문제 또는 영화, 도서 등을 추천한다. 본 논문에서는 추천 시스템의 콘텐츠 기반 알고리즘을 이용하여 추천 알고리즘 사례를 정리하고 학습자의 학습률 향상을 위한 연구를 진행한다.

II. 관련연구

1. 추천 시스템(Recommendation System)

추천 시스템(Recommendation System)이란 정보 필터링 기술의 일종으로, 사용자가 원하는 정보를 추천하는 시스템이다. 추천 시스템은 몇 가지 알고리즘의 동작 방식에 따라 연관성 규칙 분석, 협업 필터링, 콘텐츠 기반 필터링, 하이브리드로 분류된다.

먼저 연관성 규칙 분석 추천은 협업 필터링 추천으로 소비자들의 상품 클릭이나 구매 또는 후기에 대한 데이터로 상품을 추천하는 기법이다. 하지만 기존 소비자의 데이터에 기반으로 추천하기 때문에 신규 사용자 또는 신규 상품은 추천할 수 없는 콜드 스타트(cold start) 문제가 발생한다.

다음으로 콘텐츠 기반 필터링 추천은 협업 필터링과 다르게 소비자의 클릭과 구매와 후기를 고려하지 않고 상품의 특성을 고려하기 때문에 콜드 스타트(cold start) 문제는 발생하지 않는다. 이 콘텐츠 기반 필터링 추천은 본 연구에서 사용한 방식이다.

하이브리드 추천은 협업 필터링과 콘텐츠 기반 필터링 방식을 상호 보완하여 추천하는 기법이다. 콜드 스타트(cold start) 문제를 해결하기 위해 콘텐츠 기반 필터링 기술로 분석하고 충분한 데이터가 생기면 협업 필터링으로 추천 정확성을 높인다^[1].

2. Word2Vec

추천시스템의 콘텐츠 기반 필터링인 Word2Vec은 스스로 학습하여 비슷한 단어(유의어)를 찾아내는 예측 방식이며 예측 모델을 학습하면서 비슷한 단어가 비슷한 벡터로 표현된다. 모델은 2가지 방식으로 나뉘는데 CBOW와 SG(skip-gram)이다.

CBOW 모델은 주변 단어(맥락)로 타겟 단어를 예측한다. 주변 단어(맥락)의 범위를 window라고 부르며 데이터 셋을 만들 때 sliding window라는 방법을 사용한다. 이때 window는 우리의 학습 데이터가 된다.

SG(skip-gram) 모델은 타겟 단어를 이용해서 주변 단어(맥락)를 예측한다. 그러므로 크기가 큰 데이터 셋에 적합한 모델이다^[2]. 다음 그림 1은 CBOW와 SG(skip-gram) 모델 구조적 그림이다.

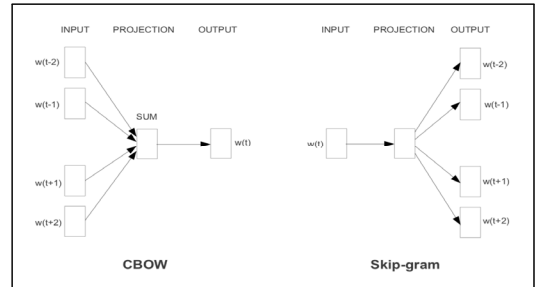


그림 1. CBOW와 skip-gram 모델 그림
Fig. 1. CBOW and skip-gram model picture.

3. 텍스트 유사도(Text similarity)

텍스트 유사도는 자연어 처리 분야 중 하나이며 텍스트 간에 유사도를 표현하는 방식이다. 관련성이 높은 문서를 탐색하여 유사도가 높은 문서를 추천해주거나 관련 키워드를 찾아주는 정보 탐색 기술이다.

유사도 측정 방법은 크게 4가지 알고리즘으로 자카드 유사도, 유클리디안 유사도, 맨하탄 유사도, 코사인 유사도로 분류된다.

다음 그림 2은 코사인 유사도 각도에 따른 값 시각화로 나타낸 그림이다.

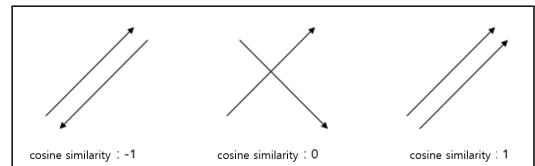


그림 2. 코사인 유사도 각도에 따른 값
Fig. 2. Value according to cosine similarity angle

본 연구에서 사용된 텍스트 유사도 기법은 코사인 유사도이며 두 벡터 간의 각도를 코사인 방식을 이용하여 유사도를 구한다. 그림 2을 보면 두 벡터의 방향이 완전히 같을 경우, 1의 값을 나타내며 완전히 반대 방향일 경우 -1의 값을 나타낸다. 두 벡터의 방향이 90°의 각도일 경우 서로 관계가 없으므로 0의 값을 나타낸다. 이때 코사인 유사도는 양수 값인 0에서 1 사이의 값을 표현한다. 그러므로 1에 가까운 값이 나올수록 유사한 텍스트임을 의미한다³⁾.

III. 연구 설계

딥러닝 기반 알고리즘인 콘텐츠 기반 필터의 Word2Vec 모델과 텍스트 유사도 중 하나인 코사인 유사도를 사용하여 문제 추천 알고리즘을 연구하여 문맥상 비슷한 위치에서의 단어들은 비슷한 뜻을 찾아내어 이를 이용한 문제 추천을 해보고자 한다.

먼저 컴퓨터 활용능력 1급 기출문제 데이터로 학습 훈련을 진행하였다. 전체적인 흐름은 그림3과 같다.

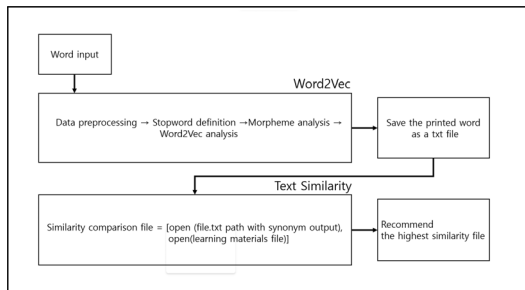


그림 3. 문제 추천 알고리즘 절차
 Fig. 3. Problem Recommendation Algorithm Procedure

알고리즘은 실행하기 전에 훈련 결과는 데이터양에도 영향을 미친다. 기출문제 1개년도의 결과와 5개년의 차이는 다음 표와 같다.

표 1. 데이터양에 따른 결과 비교
 Table 1. Comparison of results according to the amount of data.

1 year Sample questions	5 years Sample questions
[('Group', 0.2252490371465683), ('a lead', 0.19875623285770416), ('Screen', 0.15622398257255554),	[('Service', 0.995297372341156), ('send', 0.9947184324264526), ('protocol', 0.9932874441146851), ('network',

('outline', 0.14003810286521912), ('job', 0.1312871277332306), ('print', 0.1298627257347107), ('condirion', 0.12819676101207733), ('disk', 0.1269320696592331), ('Page', 0.12029766291379929), ('Service', 0.11899034678936005)]	(0.9928691387176514), ('server', 0.9921736717224121), ('Communication', 0.9919383525848389), ('packet', 0.988865077495575), ('computer', 0.9880760312080383), ('application', 0.9877139925956726), ('account', 0.9863465428352356)]
--	---

입력 단어는 모두 '인터넷'으로, 1개년일 때 나온 유의어는 '그룹', '머리글', '화면' 등 인터넷과 관련 있는 단어가 아니다. 정확도 또한 0.3 이하 수치를 보였다. 이에 비해 5개년의 유의어는 '서비스', '전송', '프로토콜' 등 '인터넷'과 관련된 단어가 나왔음을 알 수 있다. 정확도는 1과 가까워 높은 수치를 나타낸다. 본연구에선 27개년의 기출문제를 사용했으며 총 1617개의 문제로 구성했다.

표 1처럼 '인터넷'을 입력했을 때 학습하여 나오는 유의어는 정작 1과목 인터넷과 관련 없는 3과목의 데이터 베이스 단어인 '그룹', '머리글' 등의 단어가 출력되는 현상이 발생할 수 있다. 이 문제를 방지하기 위해 다음 그림 4와 같이 전체 데이터 셋을 과목에 따라 나눈다.

```

while True:
    sub = int(input(" enter the subject: "))
    if sub == 1:
        train_data = pd.read_table("number_sub1.txt")
        break
    elif sub == 2:
        train_data = pd.read_table("number_sub2.txt")
        break
    elif sub == 3:
        train_data = pd.read_table("number_sub3.txt")
        break
    elif sub == 0:
        train_data = pd.read_table("number.txt")
        break
    else:
        print("Please enter the appropriate subject ")
        continue
  
```

그림 4. 과목 데이터 셋 분류 코드
 Fig. 4. Subject Dataset Classification Code

학습자가 1을 입력했을 때 1과목만 모아둔 데이터 셋인 number_sub1 을 학습하고 외에 2와 3을 입력했을 때 경우도 같은 원리이다. 또한, 0을 입력했을 때 전체 학습 자료 즉 모든 과목을 합쳐놓은 데이터 셋을 가지고 학습한다.

다음 표는 그림 5에서 과목에 따라 분류한 데이터 셋에서 핵심 단어를 뽑아 훈련하여 나온 유의어 결과이다.

표 2. 유의어 분류 예시

Table 2. Examples of similar word classification.

Internet_(subject 1)	sheet_(subject 2)	database_(subject 3)
[(('local', 0.9956769943237305), ('wizard', 0.9953127503395081), ('control panel', 0.9925265908241272), ('port', 0.9914007186889648), ('manager', 0.9913411140441895), ('Bluetooth', 0.990932285785675), ('right', 0.9904851913452148), ('model', 0.9904519319534302), ('rest', 0.9899463653564453), ('allocation', 0.9898651242256165))]	[(('dialogbox', 0.9877369999885559), ('mouse', 0.9840195178985596), ('serialnumber', 0.9838653802871704), ('tap', 0.9829216599464417), ('screen', 0.98145592212677), ('layout', 0.9802887439727783), ('output', 0.9776901006698608), ('customization', 0.9776230454444885), ('top', 0.9751209020614624), ('row', 0.973201334476471))]	[(('Access', 0.9881507754325867), ('Excel', 0.9871151447296143), ('control', 0.9846843481063843), ('file', 0.9815048575401306), ('normalization', 0.980171263217926), ('concept', 0.9796326160430908), ('security', 0.9793514609336853), ('relay', 0.9788631796836853), ('schema', 0.9788497686386108), ('password', 0.9786430597305298))]

데이터 셋을 나눈 후 관련된 단어가 나왔음을 표2를 통해 알 수 있다. 왼쪽으로부터 1과목, 2과목, 3과목 순이다. 1과목은 인터넷과 관련된 단어가 제대로 나왔음을 알 수 있다. 엑셀과 액세스 또한 단어에 관련된 단어들이 나왔음을 알 수 있다.

학습자가 말뭉치 중 핵심 단어를 입력하면 그림 4와 같은 흐름으로 입력된 단어를 통해 Word2Vec 프로그램을 실행하면 유의어들이 출력된다. 숫자나 특수 단어와 같은 불용어를 방지하기 위해 데이터 전처리를 통해 한국어와 영어 외 다른 문자를 제거하여 분석에 적합한 구조를 만들어낸다^[4]. 이때 순수 한국어와 영어만 남아있어도 필요 없는 조사, 접속사 또는 명사 등을 제거해야 할 필요가 있다. 따라서 불용어를 얼마나 처리하느냐에 따라 출력되는 결과도 다르다.

표 3. 불용어 정의에 따른 결과 비교

Table 3. Comparison of results according to stopword definitions

stopword 100	stopword 300
[(('internet', 0.9886546730995178), ('server', 0.9815250635147095), ('give', 0.98111891746521), ('service', 0.9803304672241211), ('path', 0.97934889793396), ('layer', 0.9754973649978638), ('send', 0.9708195924758911), ('packet', 0.9689074754714966), ('mail', 0.9674486517906189), ('provision', 0.9656217694282532))]	[(('FTP', 0.9815735816955566), ('internet', 0.9802030920982361), ('access', 0.9771339297294617), ('mail', 0.9750026464646228), ('service', 0.9738290905952454), ('electronicmail', 0.9680852890014648), ('email', 0.9678438901901245), ('Anonymous', 0.9670506715774536), ('server', 0.9665138721466064), ('browser', 0.9632279276847839))]

불용어는 직접 정의할 수 있다. 따라서 필요 없는 불용어를 정의할수록 유의어의 정확도가 높아짐을 표 3을 통해 알 수 있다. 단어는 '프로토콜'로 훈련 시켰으며 불용어를 100개 정의한 것보다 300개를 정의했을 때 관련된 단어가 나왔음을 알 수 있다. 위 표는 불용어 정의를 얼마나 했는지에 따라 나온 결과 비교이다.

불용어를 300개 정의한 결과, 100개를 정의했을 때보다 프로토콜과 관련된 단어들이 나왔음을 알 수 있다.

한국어는 영어와 달리 띄어쓰기만으로 단어를 분리하기엔 조사나 어미가 있으므로 형태소에 따라 단어 분리가 필요로 한다. 본 연구에선 파이썬의 형태소 분석 패키지인 KoNLpy 의 Okt 를 이용해 정규화를 진행하였다. 정규화된 데이터 셋을 기반으로 Word2Vec 을 훈련 시키면 단어가 분리되어 유의어가 나올 수 있도록 한다^[5].

다음 그림은 텍스트 유사도 프로젝트에서 유의어와 학습 파일 간 비교를 하기 위해 쓰인 코드이다.

```
doc_list = [open('learned synonym file', 'r', encoding="cp949").readline(),]

for i in range(1, 61, 1):
    op = open('num1-60\{number}d.txt', 'r', encoding="utf-8").readline()
    doc_list.append(op)
```

그림 5. 유의어 및 학습 파일 간 비교

Fig. 5. A comparison between a significant word and a learning file.

출력된 단어 txt는 텍스트 유사도 프로젝트에 연결하여 비교할 기출문제의 유사도를 구한다. 비교할 기출문제는 총 60문제로 그림 5의 코드를 이용해 효율성과 편리성을 위해 반복문을 이용하여 유사도를 검사하였다. 유의어에 대한 문제들의 유사도 결과는 4장 실험 및 결과에서 확인하고자 한다.

IV. 실험 및 결과

Word2Vec를 학습할 때 벡터 크기(차원)와 윈도우 크기에 따라서도 정확도는 달라진다. 다음 표 4는 벡터 차원과 윈도우 크기에 따른 정확도를 나타낸다. 이때 가장 정확도가 큰 단어를 나타냈으며 모두 제 1과목의 단어인 '인터넷'을 학습한 결과이다.

확연히 보이는 정확도 수치 차이는 아니지만 윈도우 크기가 커질수록 정확도는 낮아지고 벡터의 크기가 높아질수록 정확도는 높아지는 특징을 보였다. 벡터 크기가 커짐에 따라 정확성의 차이는 크게 없었으며 이에 따라

크기 조절을 하여 최적의 정밀도를 재현할 수 있다^[6].

표 4. 벡터와 윈도우 크기에 따른 결과 비교

Table 4. Compare the results according to the vector size and window size.

	vector size 100	vector size 500
window size 3	('host', 0.9977129101753235)	('IP', 0.9994959235191345)
window size 6	('protocol', 0.9968869686126709)	('IP', 0.9991803765296936)

3장 연구설계에서 최적의 결과가 나오도록 설정하여 이 환경을 바탕으로 진행하였다. 다음 그림 6은 실제 컴퓨터 활용 자격증에서 제공된 문제이다. 이 문제를 기반으로 최적의 문제를 추천하도록 알고리즘을 진행해보고자 한다.

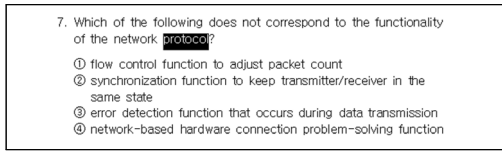


그림 6. 컴퓨터 활용 자격증 기출문제

Fig. 6. Value according to cosine similarity angle

학습자가 '프로토콜'을 핵심 단어라고 판단하여 입력 되었다고 가정했을 때 Word2Vec의 결과는 다음 표 5와 같다.

표 5. 입력 단어에 따른 유의어 결과

Table 5. Significance results according to the input word.

input word - protocol
[('internet', 0.9990586042404175), ('IP', 0.999035120010376), ('TCP/IP', 0.9989084005355835), ('packet', 0.9988957643508911), ('host', 0.9988329410552979), ('server', 0.9988203048706055), ('send', 0.9987317323684692), ('service', 0.9986860156059265), ('DNS', 0.998624861240387), ('domain', 0.9985417127609253)]

표 5에서 나온 단어는 '인터넷', 'IP', '패킷' 등 정보 통신과 관련된 단어가 나왔음을 알 수 있다. 가장 정확도가 높은 단어인 '인터넷'을 txt 파일로 저장하여 저장된 파일을 그림 5과 같이 텍스트 유사도 프로젝트와 연결한다.

텍스트 유사도에 저장된 기출문제들과 비교한 결과는 다음 그림 7과 같다.

그림 7. 유의어와 기출문제 간 유사도

Fig. 7. Similarity between synonyms and previous questions.

학습자가 입력한 단어 '프로토콜'을 학습한 결과이다. 가장 유의한 단어는 '인터넷'이며 이때 왼쪽 상단부 가장 첫 번째는 저장된 기출문제의 1번 문제이다. 총 6개의 기출문제를 활용하였으며 총 360개의 문제 중 유의어와 비교한 결과 13개의 파일에서 유사도가 나왔으며 가장 높은 유사도를 나타낸 파일은 0.25918698이다. 다음 그림 9는 높은 유사도가 나온 문제인 2019년도 1회 컴퓨터 활용 1급 필기의 10번 문제이다.

다음 문제는 정보 통신과 관련된 문제임을 알 수 있다. 그러므로 본 문제를 추천하여 학습자에게 제공한다.

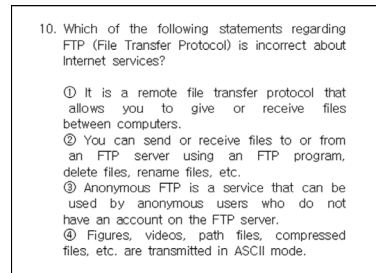


그림 8. 유사도가 가장 높은 문제

Fig. 8. The question with the highest similarity.

V. 결 론

종합적으로 정리했을 때 불용어 정의를 추가할수록, 데이터의 양이 적을수록 정확도는 낮아진다. 따라서 제한된 데이터의 양으로는 한계가 있으며 이는 대처해야 할 문제라고 생각된다.

텍스트 유사도만으로 문제 추천을 하면 문장의 의미는 고려되지 않은 채 추천될 것이다. 이를 방지하기 위한 Word2Vec 기법을 융합시켰으며 그 결과 의미까지 고려된 문제를 추천할 수 있다.

References

- [1] Soojung Lee "A Stepwise Rating Prediction Method for Recommender Systems", The Journal of the Institute of Internet, Broadcasting and Communication, Vol.21 No.4, pp.183-188, 2021
DOI:http://jiibc.iibc.kr/read.php?pageGubun=journals_earch&pageNm=article&search=Recommender&journal=%EC%A0%9C21%EA%B6%8C%20%EC%A0%9C4%ED%98%B8&code=399270&issue=399270&Page=1&year=2021&searchType=all&searchValue=Recommender
- [2] Su-Mi Shin, Kyung-Chang Kim, "Addressing the New User Problem of Recommender Systems Based on Word Embedding Learning and Skip-gram Modelling", Journal of The Korea Society of Computer and Information, Vol. 21 No. 7, pp. 9-16, July 2016.
DOI:https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06724206
- [3] Heeryong Noh, Hyunchul Ahn, "A Study on the Recommendation Algorithm based on Trust/Distrust Relationship Network Analysis", Journal of Information Technology Applications and Management, Vol. 24 No. 1, pp. 169-185, Mar 2017.
DOI:http://koreascience.or.kr/article/JAKO201731063314125.page
- [4] Jee-Uk Heu, "Korean Language Clustering using Word2Vec", The Journal of The Institute of Internet, Broadcasting and Communication (IIBC), Vol. 18, No. 5, pp.25-30, Oct 2018
DOI:http://jiibc.iibc.kr/read.php?pageGubun=journals_earch&pageNm=article&search=Word2vec&journal=%EC%A0%9C18%EA%B6%8C%20%EC%A0%9C5%ED%98%B8&code=340376&issue=340376&Page=1&year=2018&searchType=all&searchValue=Word2vec
- [5] "Visualization through Web Crawling and Morphological Analysis of National Petition Site", Proceedings of KIIT Conference, pp.353-356, Oct 2020
DOI:https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10490823
- [6] Jeong-Myeong Choi, Yu-Seop Kim, "Performance Comparison of Word Embedding Model according to Variation of Parameters in Movie Review Sentiment Analysis", Journal of Computing Science and Engineering, pp. 1400-1402, Dec 2019.
DOI:https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09301946

저 자 소 개

임 민 아(준회원)

• Min-Ah Lim is currently attending the software department at Anyang University. She research interests include Deep Learning, Big Data, Data Analysis, etc.

황 승 연(준회원)

• Seung-Yeon Hwang received his BS in Department of Computer Engineering at Tech University of Korea in 2019. He is currently studying MS in Department of Computer Engineering at Anyang University. His research interests include Deep Learning, Big Data, Data Analysis, Machine Learning, etc.

김 정 준(정회원)

• Jeong Joon Kim received his BS and MS in Computer Science at Konkuk University in 2003 and 2005, respectively. In 2010, he received his PhD in at Konkuk University. He is currently a professor at the department of Computer Science at Anyang University. His research interests include Database Systems, Big Data, Semantic Web, Geographic Information Systems (GIS) and Ubiquitous Sensor Network (USN), etc.