

자연어 처리 기반 맞춤형 트윗 추천 시스템

이현창¹, 유동필¹, 정가빈¹, 남용욱¹, 김용혁^{2*}
¹광운대학교 컴퓨터과학과 학생, ²광운대학교 컴퓨터과학과 교수

Natural Language Processing-based Personalized Twitter Recommendation System

Hyeon-Chang Lee¹, Dong-Pil Yu¹, Ga-Bin Jung¹, Yong-Wook Nam¹, Yong-Hyuk Kim^{2*}
¹Student, Dept. Comp. Sci, Kwangwoon Univ. ²Professor, Dept. Comp. Sci, Kwangwoon Univ.

요 약 트위터 사용자는 팔로우, 리트윗 등을 사용하여 자신이 관심 있어 하는 트윗을 찾는다. 하지만 사용자가 3억여 명에 달하는 트위터에서 사용자가 관심 있는 트윗을 찾기는 힘든 일이다. 이를 해결하기 위해 본 논문에서는 사용자 맞춤형 트윗 추천 시스템을 개발하였다. 우선, 사용자에게 추천할 수 있을 만한 가치가 있는 트윗을 수집하기 위해 현재 트렌드를 수집하고, 트렌드에 대해 이야기하는 인기 있는 트윗들을 수집한다. 이후 사용자를 분석하고 맞춤형 트윗을 추천하기 위해 사용자의 트윗과 수집한 트윗을 범주화한다. 최종적으로 웹서비스를 이용하여 사용자에게 본인과 카테고리가 일치하는 트윗과 관심사가 일치하는 사용자를 추천해준다. 결과적으로 67.2%로 적절한 트윗을 추천하였다.

주제어 : 머신러닝, 트위터, SNS, 자연어 처리, 키워드 추출

Abstract Twitter users use 'Following', 'Retweet' and so on to find tweets that they are interested in. However, it is difficult for users to find tweets that are of interest to them on Twitter, which has more than 300 million users. In this paper, we developed a customized tweet recommendation system to resolve it. First, we gather current trends to collect tweets that are worth recommending to users and popular tweets that talk about trends. Later, to analyze users and recommend customized tweets, the users' tweets and the collected tweets are categorized. Finally, using Web service, we recommend tweets that match with user categorization and users whose interests match. Consequentially, we recommended 67.2% of proper tweet.

Key Words : Machine learning, Twitter, SNS, Natural language processing, Keyword extracting

1. 서론

SNS(Social Network Service)는 관심사 공유, 의사소통, 인맥 확대 등을 통해 사회적 관계를 생성하고 강화해주는 온라인 플랫폼을 이야기한다[1]. 그 중 트위터(Twitter)은 한글 기준 140자 내에 자신의 의견을 게시할 수 있는 SNS로 페이스북, 인스타그램 등에 비해 익명성이 높기 때문에 사람들이 자신의 의견을 자신 있게

시하는 경향이 있다[2]. 게시한 글은 하나의 타임라인(Timeline)을 형성하며 무수한 사용자들이 동시에 의사소통하게 된다. 트위터 사용자를 팔로우(Follow)하게 되면 팔로우한 사용자의 트윗을 지속해서 확인할 수 있고 자신은 그 유저의 팔로워가 된다. 리트윗(Retweet) 기능을 이용하면 다른 사람의 트윗을 다시 트윗하여 사람들에게 알릴 수 있고 이러한 특성으로 트위터는 이전 매체와 다르게 빠르고 효율적이며, 정보 공유 및 관계 형성이

*This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2015R1D1A1A01060105).

*Corresponding Author : Yong-Hyuk Kim (ydhfly@kw.ac.kr)

Received September 13, 2018

Revised October 22, 2018

Accepted December 20, 2018

Published December 28, 2018

쉽다는 데 특징이 있다[3]. 하지만 현재 사용자가 약 3억 3천 명[4]에 달할 정도로 많은 사람이 사용하고 하루에 5억개 이상의 데이터가 생성되는 트위터 내에서 자신이 관심있어 하는 트윗을 찾는 일은 매우 힘든일이다. 자신이 팔로우 하는 계정이 인기 있는 트윗을 쓰거나[5] 리트윗을 하지 않으면 타임라인에서 발견할 수 없고, 트위터 기본 검색 시스템은 리트윗 순으로 정렬해주는 기능도 제공해주지 않아 검색 또한 쉽지 않다. 이를 해결하기 위해서 본 연구에서는 리트윗 횟수와 트랜드를 이용하여 인기 있는 트윗을 수집하였던 이전 연구[6-9]와 결합하여 사용자에게 웹을 이용해 맞춤형 서비스를 제공했다. 맞춤형 트윗 서비스란, 사용자의 관심사를 분석하고 사용자의 관심사와 일치하는 인기 있는 트윗을 추천해주는 서비스를 말한다. 이를 위해 추천해줄만한 가치가 있는 인기있는 트윗을 수집하고 수집한 트윗과 사용자의 트윗을 분석하여 분류한다. 최종적으로 웹을 이용해 사용자가 원하는 트윗을 추천한다.

본 논문은 6개의 절로 구성되어 있다. 먼저, 2절에서는 관련연구 및 기술, 3절에서는 전체적인 서비스 아키텍처를, 4절에서는 트윗을 분석하는 키워드 추출, 5절에서는 맞춤형 트윗 시스템의 웹서비스 예시를, 마지막으로 결론을 내린다.

2. 관련 연구 및 기술

2.1 트윗 수집

본 논문에서 사용자에게 추천하는 트윗은 사용자 관심 있을 만한 인기 있는 트윗이다. 인기 있는 트윗이란 트위터 사용자들에게 많이 이야기되는 트윗을 의미하며, 리트윗 수를 인기의 척도로 볼 수 있다. 이전연구[6-9]에서 트위터 트랜드를 수집하고, 이에 대한 트위터 사용자들의 의견 중 리트윗이 많이 된 트윗을 따로 DB에 저장하는 시스템을 구현하였다. 이에 더 나아가 본 논문에서는 키워드 분류를 위한 네이버 키워드 집합 또한 수집하는 기능을 추가하였다.

2.2 단어 임베딩(Word Embedding)

단어 임베딩이란 텍스트를 구성하는 단어를 수치화하여 벡터공간에 배치하는 방법의 일종이다. 단어 임베딩 방법 중 하나인 word2vec[10,11]은 단어의 앞뒤에 나오는 단어로 벡터공간에 재배치하기 때문에 그 단어의

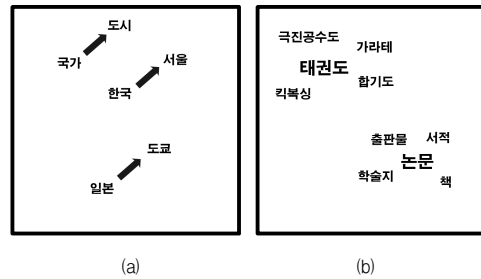


Fig. 1. Example of word2vec in the vector space.

(a) Vector offsets for three word and

(b) Example of word placement in vector space

의미를 보존하는 특징이 있다. Fig. 1의 (a)는 앞서 나온 특성을 이용한 예시로 “일본”이라는 벡터에서 “도쿄”이라는 벡터를 빼고, “서울”에 대한 벡터를 더하면 “한국”이라는 벡터를 얻을 수 있다는 것을 표현하고 있다. Fig. 1의 (b)는 word2vec의 특징을 나타내는 그림으로 벡터 공간의 단어들의 배치의 예시를 보여준다. 벡터 공간의 “태권도”라는 벡터는 “극진공수도”, “가라테”, “합기도”, “킥복싱”이라는 벡터와 가까운 거리에 있는 것을 보여준다. “논문” 또한 “출판물”, “서적”, “학술지”, “책”이라는 벡터와 거리가 짧은데, 이렇게 의미가 비슷한 단어일수록 서로 가깝게 배치한다. 즉, word2vec를 사용하면 단어의 의미를 보존해서 벡터화 시키고, 이를 이용해서 쉽게 단어 간의 유사도를 측정할 수 있다[12,13]. 본 논문에서는 word2vec의 특징인 유사도 측정을 사용하여 트윗을 분류하는데 사용하였다.

2.3 형태소 분류

형태소 분류는 어미, 조사 등이 붙어 있어 구분할 수 없는 한국어를 word2vec을 학습시키기 위해 명사, 동사, 형용사 등으로 형태소로 분석을 해서 단어화 시키는 과정을 뜻한다. 형태소 분류 패키지중 하나인 Konlpy[14,15]는 한국어 정보처리를 위한 파이썬 패키지로 총 5개의 형태소 분석기(Kkma, Komoran, Mecab, Twitter, Hannanum)로 이루어져 있다. 본 논문에서는 트윗이 구어체로 이루어져 있는 경우가 많기 때문에 구어체 분석에 성능이 뛰어난 “Twitter” 분석기를 사용하였다.

3. 추천 시스템 아키텍처

전체적인 아키텍처는 Fig. 2와 같다. 트윗 수집기를 사

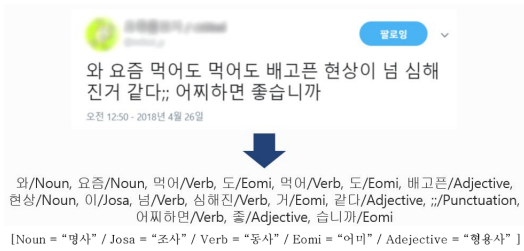


Fig. 6. Example of morphology tweet analysis

수집한 뒤 매뉴얼(Manual)하게 전처리하여서 키워드 집합을 만들었다. 트윗을 수집하면 word2vec을 이용하여 키워드집합의 키워드들과 의미유사도를 계산하고 가장 높은 의미유사도를 가진 키워드를 트윗의 키워드로 선택한다[10]. Fig. 5는 키워드 추출 과정의 전체적인 모식도를 나타낸다. 예를 들어서 트윗 수집기가 “와 요즘 먹어도 먹어도 배고픈 현상이 넘 심해 진 것 같다;; 어쩌하면 좋습니까” 라는 트윗을 수집을 하게 되면 이를 konlpy를 이용해서 형태소별로 분류하게 된다. Fig. 6과 같은 결과를 얻게 되면 분류된 단어들을 키워드집합의 모든 키워드와 비교를 하게 된다. word2vec의 ‘Similarity’는 단어 간의 의미유사도를 뜻하며 값이 높을수록 의미유사도가 높다. Table 1은 각각의 트윗의 형태소와 키워드 단어 간의 의미유사도를 word2vec을 이용해서 계산한 것을 보여준다. 위의 예시는 음식이 1.18(0.347+ 0.347 + 0.165 + 0.325)로 가장 높은 값을 가지게 되었으며 “와 요즘 먹어도 먹어도 배고픈 현상이 넘 심해 진 것 같다;; 어쩌하면 좋습니까” 라는 트윗은 “음식”이라는 키워드를 가지게 된다. word2vec 모델을 가져와서 실시간으로 대량의 데이터를 처리 하는 것은 많은 연산량을 요구하기 때문에 향상을 위해서 WM(Weight Matrix)과 TDM(Term Document Matrix)을 사용해 계산하였다. WM은 모든 키워드와 모든 단어의 유사도를 미리 구해서 행렬로 저장을 해둔 것을 말한다. TDM은 계산하려는 트윗의 모든 형태소와 단어의 빈도수를 저장한 행렬을 말한다. WM과 TDM을 행렬곱을 하게 되면 트윗의 형태소와 키워드가 곱연산이 되어서 키워드 유사도의 합이 나오게 된다. Fig. 7에서 WM의 f_1, f_2, f_3 는 키워드집합을 의미하며 v_1, v_2, v_3 는 word2vec에 저장된 모든 단어를 의미한다. TDM의 v_1, v_2, v_3 는 word2vec에 저장된 모든 단어를 뜻하며 트윗에 있는 단어가 몇 번 등장했는지 빈도수를 저장한다. WM과 TDM을 행렬 곱을 하여서 가장 높은 유사도를 가진 값을 계산하고 키워드를 추출하게 된다.

Table 1. Example of similarity comparison of tweets and keywords

Morpheme	Keyword	Similarity
“와/Noun”	“엔터테인먼트/Noun”	0.270
“요즘/Noun”	“인터넷/Noun”	0.436
“먹어/Verb”	“음식/Noun”	0.347
“도/Eomi”	“수상/Noun”	0.184
“먹어/Verb”	“음식/Noun”	0.347
“도/Eomi”	“수상/Noun”	0.184
“배고픈/Adjective”	“반려동물/Noun”	0.380
“현상/Noun”	“환경/Noun”	0.527
“이/Josa”	“이슈/Noun”	0.239
“넘/Verb”	“음악/Noun”	0.165
“심해진/Verb”	“건강/Noun”	0.270
“거/Eomi”	“음식/Noun”	0.325
“같다/Adjective”	“건축학/Noun”	0.254
“;;/Punctuation”	“컴퓨터통신/Noun”	0.172
“어쩌하면/Verb”	“컴퓨터통신/Noun”	0.247
“좋/Adjective”	“취미/Noun”	0.171
“습니까/Eomi”	“컴퓨터통신/Noun”	0.166

Table 2. Example of our keyword extraction

Tweet	Keyword
“막방 1등 고맙습니다”	“공연”
“쿠쿠쿠쿠 날씨가 더우니 힘도 없고 맛있는걸 먹어야 할거같아요”	“날씨”
“폭염이 다시 되풀이될 경우를 대비해 근본대책 마련이 필요합니다”	“재난”
“어떤 일이든 망설이지 마시고 도전하십시오”	“공부”
“한반도의 완전한 비핵화와 종전선언, 평화협정으로 가는 담대한 발걸음을 내딛을 것입니다”	“통일”

$$\begin{matrix} f_1 \\ f_2 \\ f_3 \end{matrix} \begin{bmatrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 \\ 1.1 & 0.2 & 0.1 & 0.7 & 0.8 & 0.6 & 0.4 \\ 0.8 & 0.5 & 0.8 & 1.0 & 0.8 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.5 & 0.6 & 0.2 & 0.1 & 0.7 \end{bmatrix} \times \begin{matrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{matrix} \begin{matrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{matrix} = \begin{matrix} 0.8 \\ 0.7 \\ 0.4 \end{matrix}$$

Fig. 7. Example of calculating WM and TDM

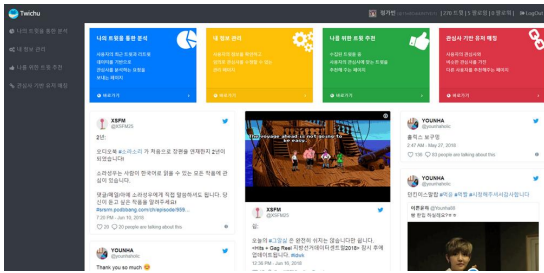


Fig. 8. Example of our Web service

4.2 검증 방법

실제 서비스를 하여서 검증하는 방법이 있지만, 실행하기가 어려워 표본 추출을 통해서 키워드가 적절한지 검사하였다. 검사방법은 트윗을 1,000 개를 뽑아서 키워드를 추출한 뒤, 5명 중 3명 이상이 적절히 키워드가 추출되었다고 생각한 것을 키워드가 적절하다고 판단하였다. 그 결과 1,000 개의 트윗 중 672개가 적절하다고 판단되었다. Table 2는 적절하다고 판단된 트윗별 키워드 예시이다.

5. 서비스 제공 웹사이트

서비스를 제공하기 위해서 웹 사이트를 제작하였고, 웹 개발 환경으로는 Mongo DB, Express JS, Angular JS, Vue.js를 사용하였다. 제공되는 서비스는 4 가지로

Table 3. List of information that can be obtained after Twitter login

Field	Type	Description
keywords	String	Interested Keywords
is_analyzing	Boolean	Check if analysis is in progress
get_tweets_count	Number	Number of Tweets
get_retweets_count	Number	Number of Retweets
id_str	String	Number ID
name	String	Name
screen_name	String	ID
location	String	Profile location
description	String	Profile description
url	String	Profile url
followers_count	Number	Number of follower
friends_count	Number	Number of friends
statuses_count	Number	Number of tweet
lang	String	Language
profile_image_url_https	String	Profile image
access_token	String	Access Token
access_token_secret	String	Access Token Secret
created_at	String	Created day



Fig. 9. Example of tweet analysis function in our Web service

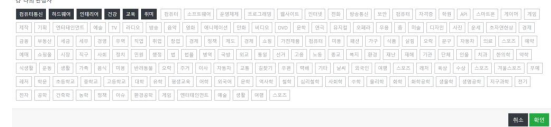


Fig. 10. Example of personal information management

사용자의 트윗을 수집해 분석해주는 “나의 트윗분석”, 키워드를 수정할 수 있는 “내정보 관리”, 사용자의 키워드에 맞게 트윗을 추천해주는 “나를 위한 트윗추천”, 관심사가 같은 유저를 추천해주는 “관심사 기반 유저매칭”이다. 웹사이트의 외관은 Fig. 8과 같다. 사용자의 트위터 정보를 가져오는 방식은 사이트에 가입하는 방식이 아닌 OAuth를 활용해서 트위터의 계정으로 로그인 하는 방식으로 가져온다. 사용자가 로그인을 하게 되면 Table 3과 같이 정보를 얻을 수 있게 된다. 먼저 사용자의 트윗과

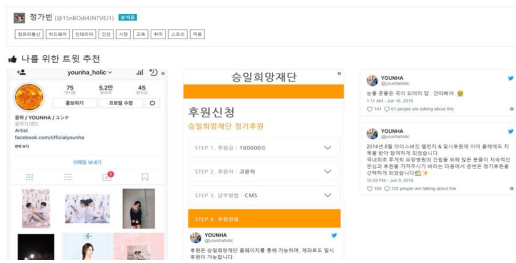


Fig. 11. Example of tweet Recommend function in our Web service

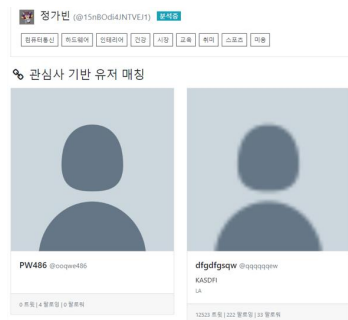


Fig. 12. User matching based on user interest

리트윗 한 트윗들을 통해 사용자가 어떤 키워드에 관심이 있어 하는지 알아낸다.

사용자는 로그인하고 자신의 트윗과 리트윗 개수를 결정한 뒤 분석을 요청하게 되면 나의 트윗을 분석해서 나에게 맞는 키워드를 Fig. 9와 같이 볼 수 있다. 사용자는 “내정보 관리” 페이지에서 자신의 트윗과 리트윗으로 분석된 관심사가 마음에 들지 않을 경우 키워드 셋에서 임의로 선택할 수 있다. Fig. 10은 “내 정보 관리 사이트”의 자신의 키워드를 확인하는 것을 보여준다. “나를 위한 트윗 추천” 페이지에서는 본 논문에서 제공하는 맞춤형 트윗을 Fig 11과 같이 추천받을 수 있다. 마지막으로 “관심사 기반 유저 매칭” 페이지에서 관심사가 같은 사용자를 추천받을 수 있다. Fig. 12을 보게 되면 “PW486”의 관심사는 “음식”, “복지”, “컴퓨터 통신”, “컴퓨터”, “하드웨어”, “소프트웨어”, “운영체제”, “프로그래밍”, “TV”이고 “dfgdfgsqw”의 관심사는 “컴퓨터 통신”, “TV”로 관심사가 일치하는 유저를 추천해주는 것을 볼 수 있다.

6. 요약 및 한계점

본 연구에서 우리는 트윗들을 수집하고 범주화하여 사용자에게 트윗을 추천하거나 사용자를 추천해주는 연구를 하였다. 인기있는 트윗을 트렌드와 리트윗수를 이용하여 수집하였고, 사용자의 트윗과 리트윗, 인기 있는 트윗을 분석하여 word2vec과 konlpy를 사용하여 범주화하였다. 이를 사용해 사용자의 관심사를 분석하고 이에 맞는 트윗을 추천해 주었다. 키워드 추출기는 67.2%로 적절히 키워드를 추출하였다. 최종적으로 웹 서비스를 이용하여 사용자에게 맞춤형 서비스를 제공하였다. 하지만 한계점이 존재한다. 우선 분석대상이 트윗이기 때문에 나무위키보다는 실제 트윗을 사용하여 학습시키는 경우, 키워드 추출에 훨씬 좋지만, API의 한계로 나무위키 데이터를 대신 사용을 하였기에 실시간으로 추가되는 트윗들의 문장을 반영하기 힘든 문제가 있다. 그리고 키워드 집합을 설정을 하는데 있어, 언어적 전문가가 없었기 때문에 임시로 네이버 지식인의 범주화를 이용을 하여서 중요한 키워드를 놓쳤을 가능성이 높다.

향후 연구에서 실제 트윗을 장기간 수집하여 좀 더 많이 확보해서 트윗을 이용해 word2vec를 학습하면, 트윗에 최적화된 단어 임베딩이 가능할 것이다. 또한 실제로

서비스를 해서 수많은 사용자가 자신의 키워드를 수정하고 선택하는 정보를 수집하여서 키워드의 검증을 하면, 실시간으로 키워드 집합을 업데이트하여서 성능을 향상시킬 수 있을 것이다.

REFERENCES

- [1] D. M. Boyd. & N. B. Ellison. (2007). Social network sites: Definition, history, and scholarship. *Journal of computer mediated Communication*, 13.1, 210-230.
- [2] J. S. Min. (2012). Study on Twitter users' political participatio. *Journal of Communication Science*, 12.2, 274-303.
- [3] S. H. Hur & K. S. Choi. (2012). A Study on characteristics and types of tweet in twitter. *Hanminjok Emunhakhoe*, 61, 455-494.
- [4] H. J. Kim. (2017. 07. 28). *Twitter users remain stuck ... Stock price plummeted by 14%*. yonhapnews. <http://goo.gl/3yjTD9>
- [5] M. W. Nho. (2012). Korea's Popular Celebrity Twitter Users and Celebrity Culture *Cybercommunication Academic Society* 29.4, 95-143.
- [6] H. Y. Cho, H. J. Kim, E. C. Lee, M. J. Lee, Y. W. Nam & Y. H. Kim. (2017) Twitter Data Collectionto Build Customized Tweet Recommendation System, *korea multimedia society* , , 254-255
- [7] Y. W. Nam & Y. H. Kim. (2016). A System of Storing Important Opinion about Twitter Trends, *Korean Institution of Information Scientists and Engineering*, 337-339.
- [8] Y. W. Nam & Y. H. Kim. (2016). Improving Twitter Search Function Using Twitter API. *Proceeding of journal of multimedia services convergent with art, humanities, and sociology* 8, 879-886.
- [9] S. J. Yang, J. W. Choi, S. H. Moon, Y. W. Jung, Y. W. Nam & Y. H. Kim. (2016). Opinion Mining Using Retweet Function of Twitter. *Proceedings of Journal of The Korean Institute of Intelligent System*, 26.1, 193-194.
- [10] T. Mikolov, K. Chen, G. Corrado & J Dean. (2013). Efficient estimation of word representations in vector space. *In Proceedings of Workshop at ICLR*.
- [11] T. Mikolov, I. Sutskever, K. Chen & GS. Corrado. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*. 3111-3119

[12] D. W. Ko & J. J. Yang. (2018). Korean Natural Language Processing and Analysis. *Korean Institution of Information Scientists and Engineering*, 2140-2142.

[13] D. W. Leem & H. Y. Jang. (2017). Keyword Extraction from Korean Wikipedia Using Word Similarity. *Proceedings of Journal of The Korean Institute of Intelligent System*, 850-852.

[14] E. L. Park & S. Z. Cho. (2014). KoNLPy: Korean natural language processing in Python. *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, 133-136

[15] E. L. Park & S. Z. Cho. (2014). KoNLPy: Python Korean NLP. goo.gl/1dPrka

이 현 창(Lee, Hyeon-Chang) [학생회원]



- 2018년 8월 : 광운대학교 컴퓨터소프트웨어학과 공학사
- 2018년 9월 ~ 현재 : 광운대학교 컴퓨터과학과 석사과정
- 관심분야 : 유전 알고리즘, 최적화 알고리즘

· E-Mail : qzeczxad@naver.com

유 동 필(Yu, Dong-Pil) [학생회원]



- 2018년 8월 : 광운대학교 컴퓨터소프트웨어학과 공학사
- 2018년 9월 ~ 현재 : 광운대학교 컴퓨터과학과 석사과정
- 관심분야 : 유전 알고리즘, 최적화 알고리즘, 음성인식

· E-Mail : yoodongphil@naver.com

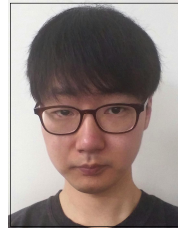
정 가 빈(Jung, Ga-Bin) [학생회원]



- 2015년 3월 ~ 현재 : 광운대학교 컴퓨터소프트웨어학과 학사과정
- 관심분야 : 데이터 베이스, 데이터 처리

· E-Mail : jgb5131@naver.com

남 용 옥(Nam, Yong-Wook) [정회원]



- 2014년 2월 : 광운대학교 컴퓨터소프트웨어학과 공학사
- 2014년 3월 ~ 현재 : 광운대학교 컴퓨터과학과 석박통합과정
- 관심분야 : 자동 작곡, 최적화 알고리즘

· E-Mail : mitssi@kw.ac.kr

김 용 혁(Kim, Yong-Hyuk) [정회원]



- 1999년 2월 : 서울대학교 전산과학 전공 이학사
- 2001년 2월 : 서울대학교 컴퓨터공학부 공학석사
- 2005년 2월 : 서울대학교 컴퓨터공학부 공학박사

· 2005년 3월 ~ 2007년 2월 : 서울대학교 반도체 공동연구소 연구원

· 2007년 3월 ~ 2012년 2월 : 광운대학교 컴퓨터소프트웨어학과 조교수

· 2012년 3월 ~ 2017년 2월 : 광운대학교 컴퓨터소프트웨어학과 부교수

· 2017년 3월 ~ 현재 : 광운대학교 소프트웨어학부 교수

· 관심분야 : 유전 알고리즘, 최적화 알고리즘

· E-Mail : yhdfly@kw.ac.kr