지도학습 머신러닝 기반 카테고리 목록 분류 및 추천 시스템 구현☆

Development of Supervised Machine Learning based Catalog Entry Classification and Recommendation System

이 형 우1* Hyung-Woo Lee

요 으

200 만명 이상의 회원을 보유하고 있는 "도매꾹" B2B 온라인 쇼핑몰인 경우70% 이상의 시장 점유율로 하루에 80만개 이상의 아이템이 판매되고 있다. 하지만, 동일하거나 유사한 물품이 서로 다른 카탈로그 엔트리에 저장 및 등록되어 있기 때문에 구매자가 아이템을 검색하는 과정에서 어려움을 느끼며 B2B 대형 쇼핑몰 관리에도 문제점이 발생하고 있다. 따라서 이에 대한 해결 방안으로 본 연구에서는 대단위 쇼핑몰 구매 정보를 기반으로 지도-학습 머신러닝 기법을 적용하여 상품에 대한 카탈로그 목록 자동 분류 및 추천 시스템을 개발하였다. 구체적으로 판매자가 자연어 형태로 물품 등록 정보를 입력하면 KONLPy 형태소 분석 과정을 수행하였으며, Naïve Bayes 분류 방식을 응용하여 물품에 가장 적합한 카탈로그 정보를 자동으로 추천해주는 시스템을 구현하였다. 정확도가 향상된 카테고리 목록을 구축하여 결과적으로 검색 속도와 쇼핑몰 매출을 향상시키는 효과가 있었다.

□ 주제어 : 카테고리 자동 추천, 온라인 B2B 소핑몰, 지도학습 기반 머신 러닝, 형태소 분석, Noïve Boyes 분류.

ABSTRACT

In the case of *Domeggook* B2B online shopping malls, it has a market share of over 70% with more than 2 million members and 800,000 items are sold per one day. However, since the same or similar items are stored and registered in different catalog entries, it is difficult for the buyer to search for items, and problems are also encountered in managing B2B large shopping malls. Therefore, in this study, we developed a catalog entry auto classification and recommendation system for products by using semi-supervised machine learning method based on previous huge shopping mall purchase information. Specifically, when the seller enters the item registration information in the form of natural language, KoNLPy morphological analysis process is performed, and the Naïve Bayes classification method is applied to implement a system that automatically recommends the most suitable catalog information for the article. As a result, it was possible to improve both the search speed and total sales of shopping mall by building accuracy in catalog entry efficiently.

region keyword: Catalog Entry Auto-Recommendation, Online B2B Shopping Mall, Supervised Machine Learning, Morphological Analysis, Naive Bayes Classification Algorithm.

1. Introduction

In addition to the development of the Internet, the use of online shopping malls is very high in B2B transactions. Especially wholesale site of GNG Commerce Co., Ltd[1]

which applies the result of this study is for a B2B on-line shopping mall with over 2 million registered members. However, previous B2B web site provide a interface in which users select and register catalog entry classification for product manually. As a result, existing passive systems of previous online shopping malls do not provide convenience to users (B2B buyer), and same and similar items are registered on different catalog in appropriately. Therefore, we want to develop an automated catalog recommendation system that solves these problems and improves the accuracy of the products catalog, the efficiency of search goods in the online shopping mall, and the increase in the total sales of B2B mall as a result.

¹ Div. of Computer Engineering, Hanshin University, Yangsan-dong, Osan, Gyeong-gi, 18101, Korea.

^{*} Corresponding author (hwlee@hs.ac.kr)
[Received 30 August 2018, Reviewed 24 October 2018, Accepted 01
December 2018]

[☆] This work was supported by the Basic Science Research Program (NRF-2017R1D1B03035040) and partly by the Technology Development Program of MSS [\$2504920].

[☆] A preliminary version of this paper was presented at APIC-IST 2018 seminar and was selected as an outstanding paper.

In this paper, we present a brief introduction to supervised machine learning for system development based on python [2] and suggest the process of analyzing and extracting meaningful information from item name entered by the user (B2B seller) using a morpheme analyzer with KoNLPy [3] using Natural Language Processing as a "semi-supervised machine learning." We propose the process of classifying categories by learning data using Naïve Bayes classification [4, 5] function. In addition, preprocessing method applied directly to the actual B2B online shopping mall site and the structure of automatic classification algorithm based on *Naive Bayes* are designed in detail and its experimental results are presented respectively.

2. Machine Learning-based catalog entry Auto-Recommendation

2.1 catalog Entry Management and its Problems

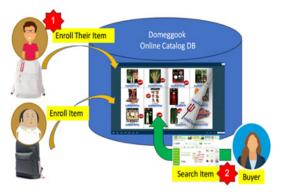
Currently, B2B large-scale online shopping mall such as GNG commerce Co., Ltd. is manually performing the process of manually selecting and registering categories as shown in the following figure. Domeggook.com site on GNG Commerce Co., Ltd is a No. 1 B2B on-line wholesale shopping mall in Rep. of Korea (On-line wholesale market share : 70%). There are over 2 million of registered member with 800,000 items sold per one day.





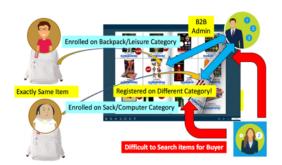
(Figure 1) Domeggook B2B Shopping Mall & its
Passive Catalog Entry Registration
Process

On Domeggook shopping mall, each seller enrolls their own item for sale. Then, buyers search for items. In detail, each seller's items are enrolled on Domeggook's online catalog DB. Then, Buyers search for items based on this catalog.



(Figure 2) Existing Item Enrollment Process on Domeggook B2B Shopping Mall from B2B Seller & Search Item Process for B2B Buyer

However, similar items are enrolled on different catalog entry. Therefore, it is difficult to manage catalog system by Domeggook administrator. For example, exactly same items were enrolled on different catalog entry (Backpack/Leisure and Sack/Computer catalog entry as follow figure). So, it is also difficult to search appropriate item by each buyer.



(Figure 3) Existing Problems on Item Enrollment and Search Process on Domeggook B2B Shopping Mall

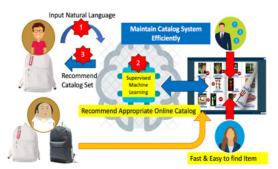
In the case of the existing system [7, 8, 9, 10], since the process is a passive catalog entry selection method, the registered product may be enrolled in an inappropriate catalog entry and may be re-registered with a previously registered catalog entry. Therefore, it is necessary to develop an advanced system for classifying and recommending the catalog entry most suitable for a product name based on a list of existing categories to be added in the process of registering into item catalog entry DB of online shopping mall by a B2B merchant (seller).

2.2 catalog Entry Recommendation with SML

As a solution, we developed catalog entry auto-classification and recommendation system using Supervised Machine-Learning(SML) [6, 7]. If a seller(B2B merchant) input item's name as a natural language format, then proposed system automatically recommends appropriate catalog entry set for appropriate enrollment on item DB.

If catalog entry auto-recommendation mechanism is implemented, same and similar item can be registered on exact and appropriate catalog (Advantage for seller). And then, it is much more fast and easy to find wanted items from wholesale on-line shopping mall (Advantage for buyer). Additionally, it is possible for the B2B admin to maintain catalog system efficiently (Advantage for B2B site administrator).

For this purpose, it is necessary to directly pre-process the product(item) names related to existing catalog entry on de



(Figure 4) Suggested Catalog Entry Recommendation System on Wholesale B2B Shopping Mall

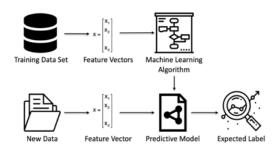
veloping a system for classification and recommendation. To overcome the problems of catalog entry classification and recommendation in large wholesale online shopping mall, we use machine learning method to efficiently classify categories based on existing registered data and automatically recommend the most suitable catalog entry. Therefore, we developed an efficient catalog entry automatic recommendation system using supervised machine learning model specifically.

3. Natural Language Processing and Morphological Analysis for SML

3.1 Semi-Supervised Machine Learning Model

Machine learning is divided into Supervised Learning, Unsupervised Learning, and Reinforcement Learning models. In order to apply the machine learning method to the automatic recommendation of the catalog entry information about the large items sold in the B2B online shopping mall, the machine learning method using the supervised learning model is applied in this study.

The supervised learning model is mainly used for prediction. Basically, it is a method of learning by using learning data using machine learning algorithm and predicting new data based on it. Semi-supervised learning is a class of machine learning tasks and techniques that also make use of unlabeled data for training typically a small amount of labeled data with a large amount of unlabeled data. Semi-su-



(Figure 5) Supervised Machine Learning Model

pervised learning [11, 12, 13, 14] falls between unsupervised learning (without any labeled training data and supervised learning (with completely labeled training data). The acquisition of labeled data for a learning problem often requires a skilled human agent. The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning.

3.2 Morphological Analysis on Catalog Entry

The user performs a process of registering the article to be registered in the online shopping mall in the name of a natural language type article. Therefore, for the items registered in the large online shopping mall, a natural language processing process is required as a preprocessing process. Natural language processing generally refers to a set of techniques for analyzing, extracting, and understanding meaningful information in text. n case of English words in natural language processing, we can divide morphemes based on white space and perform additional analysis easily by performing exception processing only for some limited cases such as "Can't" or "Didn't".

However, in the case of the Korean language, it is more difficult to process the English sentence than the natural language because the search and the mother language are additionally attached. Therefore, for the accurate and efficient processing of Korean information, natural language processing was performed through morphological analysis of parts by using KoNLPy's Twitter analyzer, a collection of Python packages[3].



(Figure 6) Example of Korean Morpheme Analysis

4. KoNLPy and Naïve Bayes Classifier

4.1 KoNLPy(3) and Twitter Analyzer

KoNLPy has adopted the GPL v3 license as open source software. Also, as the name implies, it is a package using Python and contains various analyzers (Hannanum, Kkma, Komoran, Mecab, Twitter) inside the software. Therefore, more accurate morphological analysis result can be obtained by applying the analyzer according to the characteristics of input values.

One of the analyzers included in the KoNLPy described above, the Twitter analyzer provides precise identification of parts of speech. Since there is a characteristic that the product information is additionally written in the case of the article information registered in the large-capacity online shopping mall, the remaining parts are removed using the Twitter analyzer provided by KoNLPy except for the noun and English. Through this, it was possible to extract noun information that is the core of the product name.

Hannanum	Kkma	Komoran	Mecab	Twitter
나/N	나/NP	나/NP	나/NP	나/Noun
는/기	는/JX	는/JX	는/JX	는/Josa
밥/N	밥/NNG	밥/NNG	밥/NNG	밥/Noun
을/J	을/JKO	을/JKO	을/JKO	을/Josa
먹 / P	먹/W	먹/W	먹/W	먹는 / Verb
는다/E	는/EPT	는다/EC	는다/EC	다/Eomi
	다/EFN			

(Figure 7) Examples of Natural Language Analyzers included in KoNLPy and Twitter

4.2 Naive Bayes Classifier(4)

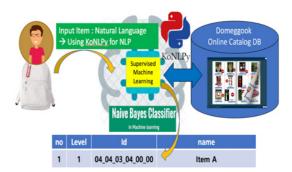
It is one of the machine learning classification algorithms

and we can use it for classification of documents or keyword search. It is a probabilistic classification that applies Bayes Theorem to the conditional probability and classifies the input data assuming independence of the probability that each element constituting the document or data. Naive Bayes should be preceded by consequent learning previously through supervised learning data as it can be implemented as a machine learning algorithm.

In this study, we applied the classification method to the classification using the Naive Bayes classifier and improved the accuracy of the classification and recommendation results regardless of the number of data included in each catalog entry.

$$egin{aligned} \log p(C_k|\mathbf{x}) & \propto \log \Bigg(p(C_k) \prod_{i=1}^n {p_{ki}}^{x_i} \Bigg) \ &= \log p(C_k) + \sum_{i=1}^n x_i \cdot \log p_{ki} \ &= b + \mathbf{w}_k^ op \mathbf{x} \end{aligned}$$

When a seller want to enroll item into Domeggook online catalog entry DB, he/she input item name as a natural language(especially Korean Language). Therefore, we used both KoNLPy/Twitter and Naïve Bayes classifier to suggest appropriate catalog entry using supervised machine learning process as follow figure.



(Figure 8) Catalog Entry Auto-Recommendation Process using both KoNLPy/Twitter and Naïve Bayes Classifier with Supervised Machine Learning Process

5. Proposed Catalog Entry Auto Recommendation System

5.1 Structure of Catalog Entry Classification

Based on the above, we propose the process applied to the automatic classification and recommendation system of product categories developed for wholesale online B2B shopping mall.



(Figure 9) Proposed Catalog Entry Auto-Recommendation System for Wholesale Online B2B Shopping Mall

First, the product catalog entry classification system is used as learning data by using the product name of each catalog entry stored through the connection between the server on which the classification system operates and the DB server. At this time, the user inputs the product name (text information) data in the process of input & register the product intends to sell in the online shopping mall. In addition, the completed information is stored in the DB server in categories, words (noun, English), number of occurrences and scores as a Predictive Model.

In order to perform accurate and efficient learning, continuous learning process is required for as many data as possible. Therefore, if a method of updating the contents stored in the DB server through a periodic learning process rather than a real-time processing method is used, it is possible to construct an efficient catalog entry recommendation system for a large-capacity online shopping mall. As a result, we implemented the following periodic learning method.

(Table 1) Predictive DB for Enrolled Items

itemCategory	wordName	wordCount	wordScore
01_01_01_01_01	@None	0	-8.750049
01_01_01_01_01	나비넥타이	1	-8.056902
01_01_01_01_01	레드	1	-8.056902
01_01_01_01_01	중	1	-8.056902
01_01_01_02_00_00	@None	0	-8.773849
01_01_01_02_00_00	KC	5	-6.982089
01_01_01_02_00_00	L	1	-8.080702
01_01_01_02_00_00	OPP	3	-7.387555
01_01_01_02_00_00	SG	1	-8.080702
01_01_01_02_00_00	X	1	-8.080702
01_01_01_02_00_00	가제	1	-8.080702
01_01_01_02_00_00	개별	5	-6.982089
01_01_01_02_00_00	고급	4	-7.164411
01_01_01_02_00_00	국내	1	-8.080702



(Figure 10) Periodic Prediction Learning Structure

When using the registered product for catalog entry fore-casting and recommendation purposes, it is implemented to minimize the load of the entire system by querying the DB server only for the data necessary for judgment. Also, in the process of text preprocessing, the filtering function was applied before learning to remove meaningless words in advance, and the accuracy of catalog entry recommendation was improved by this process.



(Figure 11) Catalog Entry Classification System Structure

5.2 Experimental Data

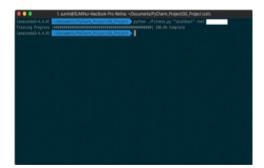
We used 6,394 data set previously stored on B2B shopping mall site of GNG Commerce Co., Ltd. The data below is part of the products list that are currently registered in the wholesale site, and our supervised machine learning system used this data set for classifying and recommending input item automatically.

(Table 2) Example of Catalog Entry Classification Learning Data

	itemCategory		
(불러/돌답레용/개업선물/보온병/덤벨	04_04_04_06_00_00		
선택가능/어린이날선물/여린이단쇄선물/	03_04_06_01_00_00		
도/건강배드/인상마스/홍상마스/인상	05_02_03_09_00_00		
이스] 고급열티ভ케이스(大) 7cmX9cm	03_05_01_08_01_00		
3종머리띠 물결모양 철사모양 스프링	01_01_02_01_99_00		
한력 고급 캐시미고무를 팔찌 급창	01_01_02_04_00_00		
P박목막린A 최고급월리티	01_01_02_03_01_00		
스틱엑스 파랑 1호 (480x345x345)-10	06_04_01_08_00_00		
8없는 굵은검정 교무를 머리끈 헤어밴	01_01_02_04_00_00		
가로x세로+날개) HDPE 링크 1-1호	06_04_02_02_00_00		
가로x세로+날개) HDPE 원크 1호	06_04_02_02_00_00		
가로x세로+날개) HDPE 핑크 2호	06_04_02_02_00_00		
가로x세로+날개) HDPE 원크 2-1호	06_04_02_02_00_00		
가로x세로+날개) HDPE 링크 4호	06_04_02_02_00_00		
가로x세로+날개) HDPE 원크 5호	06_04_02_02_00_00		
4 35 + 4 (가로x세로+날개) HDPE 은색	06 04 02 02 00 00		
o종 당성애탈시계/파선시계/당자시계/당	01 02 07 02 00 00		
가로x세로+날개) HDPE 보라 2호	06 04 02 02 00 00		
가로x세로+날개) HDPE 보라 3호	06_04_02_02_00_00		
가로x세로+날개) HDPE 보라 4호	06_04_02_02_00_00		
E드폰줄/엔드폰악세사리/갤럭시노트홍	06_03_03_03_00_00		
) + 4 (가로x세로+남개) 신형 플리백 4	06_04_02_05_00_00		
+ 4 (가로x세로+날개) 신형 폴리백 2	06_04_02_05_00_00		
30 + 4 (가로x세로+남개) 신형 플리벡	06_04_02_05_00_00		
15 + 4 (가로x세로+날개) 접착 OPP 2호	06_04_02_07_00_00		
가로x세로+날개) HDPE 검정 2-1호	06_04_02_02_00_00		
TIELLIE LUNC TOTAL AND	00 04 02 02 00 00		
,	1로x세로+날개) HDPE 검정 2−1호	HEXAER+1/10) HDPE 2/8 2-18 06_04_02_02_00_00	(紀) 세요 나가

5.3 Local Test

Local testing conducted a machine learning process for the catalog entry by performing the supervised learning process in the CLI environment through the console interface. The figure below shows the data learning screen. The learning process is performed by reading the Training Data Set from the DB and storing the learned Predictive Model in the DB again.



(Figure 12) Machine Learning Process

In case of enter the new Item, we can see recommended categories as shown in below Figure using Predictive Model and it also provide the filtering result for a search keyword.



(Figure 13) Prediction, Recommendation, and Filtering Process after Machine Learning Step

In this paper, we propose a preprocessing process for the data because we use the machine learning model based on the big data manually registered by the users. Data can be entered by users in the form of words and phrases they want. In this process, users can use the filtering function in the registration process because they include vendors, discount information, etc. in addition to descriptions of the products they register. Here, we show it based on the appearance in the CLI environment, and the original file is set as a * .txt file to separate words with a comma (,) like CSV.

```
● ● ● ・ LauringのAnnu-Mechan Pro-Reina - Occurrent PyCham (Project Carlo) Project Carlo) Proje
```

(Figure 14) Catalog Entry Accuracy Measurement

The accuracy of catalog entry discrimination and recommendation is measured as follows. First, it is confirmed that the catalog entry classification accuracy is about 37% when it is judged that it is a correct classification only when the catalog entry of the stored data is exactly 100%. This is be-

cause there are similar categories in the catalog entry classification and there are many cases belonging to the overlapping catalog entry. As a result, it is confirmed that the catalog entry is recommended with higher accuracy than the passive method. In addition, the present invention provides a function of recommending the top three categories having high relevance as well as the catalog entry corresponding to the first catalog entry in the product catalog entry classification and recommendation system, so that users can finally select appropriate catalog entry.

5.4 Field Test

In fact, we conducted a direct test process to check whether the catalog entry classification and recommendation functions were performed well in cooperation with the database of GNG Commerce, Ltd. The test results showed that we automatically recommend the top three categories with high relevance and relevance to the input keywords, as confirmed by our local tests.

In case of enter "folding seat computer desk (Korean: 접이식 좌석 컴퓨터 책상)" as the result of the web execution below, you can automatically check the three categories of "computer desk (컴퓨터 책상)", "sitting desk (좌식 책상)" and "other (기타)" automatically.



(Figure 15) Catalog Entry Recommendation Results for Keyword "folding desk computer desk"

In addition, we confirmed that "Backpack (백팩)", "Cross Bag (크로스백)" and "Travel Accessory Case (여행소품케이스)" were automatically recommended as categories when we requested registration as "Military Backpack Military Bag Traveling Bag Climbing Bag (Korean: 밀리터리 백팩 군인

가방 여행용가방 등산가방)". And its accuracy of recommendation was also acceptably high.



(Figure 16) Catalog Entry Recommendation Results for Keyword "Military Backpack Military Bag Traveling Bag Climbing Bag"

Finally, even if you input the words sep과rated by '/' such as "Golden long sword / Golden knife / Toy knife / Long knife / Ninja knife / General knife (Korean: 황금장검 / 황금칼 / 장난감칼 / 긴칼 / 닌자칼 / 장군칼)", three categories such as "Toy gun (장난감총)", "Knife / cutting / cutting tool (칼 / 도마 / 커팅기구)" and "tool play (공구놀이)" categories were automatically recommended. It was somewhat less accurate, but we could confirm that the relative catalog entry "sword (검)", "knife (칼)" and the catalog entry containing the word "toy (장난감)" were automatically recommended.



(Figure 17) Catalog Entry Recommendation Results for Keyword "Golden long sword / Golden knife / Toy knife / Long knife / Ninja knife / General knife"

As you can see in the above example, the catalog entry is recommended automatically and fairly accurately according to the input keyword. Experimental results show that the knowledge - based machine learning method can improve the classification accuracy of items and the relevance to recommended categories than manual passive catalog entry registration method. As a result, sales of B2B online shopping malls can be expected to improve.

6. Conclusion

We developed a catalog entry classification and automatic recommendation algorithm for registered products based on supervised machine learning technology for purchase information with a morphological analyzer (KoNLPy) to extract meaningful information from the product name (text information) entered by the users of large scale online B2B shopping mall. As a result, it was possible to improve the accuracy of the product more than when the shop user manually registered the shopping mall. In order to provide more accurate catalog entry classification and automatic recommendation function, it is necessary to improve the quality of learning data by eliminating redundant words in the filtering process and to correct errors in the registered catalog entry.

References

- [1] GNG Commerce B2B Online Shopping Mall Web site: http://domeggook.com
- [2] Willi Richert, Luis Pedro Coelho, "Building Machine Learning Systems with Python," Packt, 2013.
- [3] Eunjeong L. Park, Sungzoon Cho, "KoNLPy: Korean natual language processing in Python", 26th Hangul and Korean Information Processing Proceeding, 2014.
- [4] Naïve Bayes classifier : https://en.wikipedia.org/wiki/Naive_Bayes_classifier
- [5] Ki-Duk Kwon, "A Naive Bayesian-based Model of the Opponent's Policy for Efficient Multiagent Reinforcement Learning," Journal of Internet Computing and Services, vol. 9, no. 6, pp. 165-178, 2008.
- [6] Scikit-Learn: http://scikit-learn.org
- [7] Jun Lee, and YongJin KWON, "A Proposal of Methods for Extracting Temporal Information of

History-related Web Document based on Historical Objects Using Machine Learning Techniques," Journal of Internet Computing and Services, vol. 16, no. 4, pp. 39-50, 2015.

https://doi.org/10.7472/jksii.2015.16.4.39

- [8] Im San-Song, Na Cheol-Hun, and Jung Hoe-Kyung, "Design and implementation of integrated e-catalog system based on web services," Journal of Internet Computing and Services, vol. 6, no. 2, pp. 153-0, 2005.
- [9] Young-Won Seo, and Chang-soo Kim, "Clustering Algorithm using the DFP-Tree based on the MapReduce," Journal of Internet Computing and Services, vol. 16, no. 6, pp. 23-30, 2015. https://doi.org/10.7472/jksii.2015.16.6.23
- [10] Seung-Su Chun, "A study on integrating and discovery of semantic based knowledge model," Journal of Internet Computing and Services, vol. 15, no. 6, pp.

- 99-106, 2014. https://doi.org/10.7472/jksii.2014.15.6.99
- [11] V. Jothi Prakash, Dr. L.M. Nithya, "A Survey On Semi-Supervised Learning Techniques," International Journal of Computer Trends and Technology (IJCTT), Vol.8, No.1, Feb 2014. https://doi.org/10.14445/22312803/IJCTT-V8P105
- [12] Alejandro Cholaquidis, Ricardo Fraiman and Mariela Sued, "Semi-supervised learning: When and why it works," May, 2018, https://arxiv.org/pdf/1805.09180 [DOI: arXiv:1805.09180].
- [13] Semi-supervised learning : https://en.wikipedia.org/wiki/Semi-supervised_learning
- [14] Huimin Pei, Kuaini Wang, Qiang Lin, Ping Zhong, "Robust semi-supervised extreme learning machine," Knowledge-Based Systems, Vol. 159, No. 1, pp.203-220, Nov, 2018. https://doi.org/10.1016/j.knosys.2018.06.029



이 형 우(Hyung-Woo Lee)

1994년 고려대학교 컴퓨터학과(이학사) 1996년 고려대학교 대학원 컴퓨터학과(이학석사)

1999년 고려대학교 대학원 컴퓨터학과(이학박사)

2003년~현재 한신대학교 컴퓨터공학부 교수

관심분야 : 모바일/네트워크 보안, 디지털포렌식, 정보보호, etc.

E-mail: hwlee@hs.ac.kr