Journal of the Korea Convergence Society
Vol. 12. No. 9, pp. 39-47, 2021

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2021.12.9.039

A Study on the Development of Product Planning Prediction Model Using Logistic Regression Algorithm

Yeong-Hwil Ahn[1], Koo-Rack Park[2*], Dong-Hyun Kim[3], Do-Yeon Kim[1]
[1]Ph. D. Student, Division of Computer Engineering, Kongju National University
[2]Professor, Division of Computer Science & Engineering, Kongju National University
[3]Professor, Dept. of IT Artificial Intelligence, Korea Nazarene University

로지스틱 회귀 알고리즘을 활용한 상품 기획 예측 모형 개발에 관한 연구

안영휘[1], 박구락[2*], 김동현[3], 김도연[1]
[1]공주대학교 대학원 박사수료, [2]공주대학교 컴퓨터공학부 교수, [3]나사렛대학교 IT인공지능학부 교수

Abstract This study was conducted to propose a product planning prediction model using logistic regression algorithm to predict seasonal factors and rapidly changing product trends. First, we collected unstructured data of consumers in portal sites and online markets using web crawling, and analyzed meaningful information about products through preprocessing for transformation of standardized data. The datasets of 11,200 were analyzed by Logistic Regression to analyze consumer satisfaction, frequency analysis, and advantages and disadvantages of products. The result of analysis showed that the satisfaction of consumers was 92% and the defective issues of products were confirmed through frequency analysis. The results of analysis on the use satisfaction, system efficiency, and system effectiveness items of the developed product planning prediction program showed that the satisfaction was high. Defective issues are very meaningful data in that they provide information necessary for quickly recognizing the current problem of products and establishing improvement strategies.

Key Words : Big Data, Product Planning, Sentiment Analysis, Logistic Regression, Web Crawling

요 약 본 연구에서는 계절적인 요인과 급변하는 상품의 트렌드를 사전예측하기 위해 로지스틱 회귀 알고리즘을 이용한 상품기획 예측 모형을 제안하고자 수행되었다. 먼저 웹크롤링을 이용하여 포털 사이트 및 온라인 마켓의 소비자의 비정형 데이터를 수집하고 정형 데이터 변환을 위한 전처리 작업을 통해 상품에 대한 의미 있는 정보를 분석하였다. 최종 수집된 11,200개의 데이터셋은 Logistic Regression을 이용하여 상품에 대한 소비자의 만족도, 빈도분석, 상품에 대한 장점과 단점을 분석할 수 있었다. 분석 결과 소비자의 만족도는 92%이었으며, 빈도분석을 통해 상품에 대한 불량이슈를 확인할 수 있었다. 또한, 개발된 상품 기획 예측 프로그램에 대한 사용 만족도, 시스템 효율성, 시스템 효과성 항목에 대한 분석결과에서도 만족도가 높게 나타났다. 특히, 불량이슈는 상품에 대한 현 문제를 신속히 인지하고 개선 전략을 수립하는데 필요한 정보를 제공한다는 점에서 매우 의미 있는 자료가 된다.

주제어 : 빅데이터, 상품기획, 감성분석, 로직스틱 회귀분석, 웹크롤링

*Corresponding Author : Koo-Rack Park(ecgrpark@kongju.ac.kr)
Received August 31, 2021 Revised September 11, 2021
Accepted September 20, 2021 Published September 28, 2021

1. Introduction

In order to conquer the rapidly-changing on/offline new product market, all the companies including large, medium, and small companies are fiercely engaged in heated competitions to get the initiative in the market, based on their own differentiated product development process strategies. A successful new product could bring huge profits to a company[1], and the process for developing a new product is a core element having lots of effects on the successful new product[2], so that the development of new products could be a very important matter for a company to survive in the fierce competitions[3]. Even though the development of new products is an essential part for companies to have the competitiveness, it is actually very tough to succeed in new products due to the complexity of product development process. In various researches, the researchers are emphasizing the difficulty of successful development of new products[4], and it would be such a difficult task for product developers and researchers of companies to find a method to be equipped with the innovative process for the successful performance of product development[5]. As the companies that aim to develop/release new products need to reflect consumers' opinions and needs into their products, the customer needs analysis is applied in priority[6]. Thus, it would be necessary to perceive that the customers' needs are directly/indirectly involved in the corporate development of new products, and to create the product value by including various solutions to meet their needs[7]. However, it is very rare to see that the product development process is smoothly carried out with no trials and errors. In the process of producing new products, many difficulties are frequently occurring contrary to the expectation. In order to efficiently cope with such difficulties, it would be required to have the procedure for reflecting consumers' needs by introducing and applying the recent Opinion Mining, Sentiment Analysis, and Logistic Regression to the product development process. This presents the direction of rapidly-changing consumers' needs and economic feasibility of time and monetary investment by reflecting the consumers' sentimental needs in the results of analyzing the massive unstructured data of honest and plain customers' emotion, preference, and feelings, and the prediction rate through logistic regression, to the product development process. As a representative logistic regression, the supervised learning shows the 95% of examples related to the whole logistic regression. The training data which is a required element of supervised learning is a concept used for logistic regression model, and also explains the advisable corresponding relation between observed value and result value that should be provided by logistic regression. By using this corresponding relation, it is possible to predict the quantity of sales/demand, to classify a certain situation like machine failure or customer purchase/breakaway, and to predict the possibility of relevant situation. In case when selecting a model in supervised learning, it is to mapping a situation to be predicted from a certain data, or to select a method of explaining abstract patterns. The simplest type includes the linear regression analysis or Scorecards. A bit-more developed type includes Decision Trees and Neural Network. And the most developed type for now includes Ensemble Model and Deep Neural Network. This study suggests the Artificial Intelligence(AI) product recommendation model based on the linear regression and logistic regression models to predict a product with the highest purchase possibility. Which is used for the newest marketing technique for the next best action or product recommendation, by analyzing the massive unstructured consumer data of portal sites and online markets, positiveness/negativeness,

and predicting, analyzing the causes. Through the proposed model, it would be possible to improve the customer satisfaction, to proceed the corporate decision-making, and also to improve the success probability in the development of new products based on the prediction of demand for the relevant product in a certain price range.

2. Related Works

2.1 Sentiment Analysis

As a process of sentimentally analyzing the product reviews, the following Fig. 1 distinguishes the sentiments shown in texts, and understands/classifies consumers' sentiment[8].

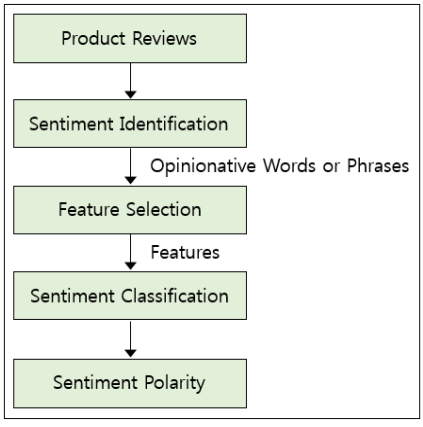

Fig. 1. Sentiment Analysis Process

The sentiment analysis which is also used as opinion mining, is the technique of analyzing writers' sentiment through unstructured data such as document or sentence with the text mining technique. The sentiment analysis based on the natural language processing started from the classification task in document level[9], developed to the sentiment analysis in sentence level[10], and even reached the level of processing the sentiment analysis in phrase level[11]. Like this, the sentiment analysis which is the natural language processing for cleansing and analyzing subjective data such as people's attitude, tendency, and opinions shown in texts, is composed of three procedures as follows. First, as the data collection step, it is to extract texts from unstructured big data. Second, as the stopword processing step, after separating into sentence unit for sentiment analysis, the words unrelated to sentiment analysis are processed as stopwords. Third, as the analysis step, after the polarity analysis for judging if the data is "positive" or "negative", the weighted value could be given by digitizing the positiveness/negativeness using the statistical technique, or the comprehensive evaluation data could be produced through the frequency analysis. composed of three procedures as follows.

First, as the data collection step, it is to extract texts from unstructured big data. Second, as the stopword processing step, after separating into sentence unit for sentiment analysis, the words unrelated to sentiment analysis are processed as stopwords. Third, as the analysis step, after the polarity analysis for judging if the data is "positive" or "negative", the weighted value could be given by digitizing the positiveness/negativeness using the statistical technique, or the comprehensive evaluation data could be produced through the frequency analysis.

2.2 Web Crawler

The web crawling means the collection of URL information and the analysis of webpages by making the rounds of web servers in the automated method by using a crawler, and also means the technique of collecting the information of web. As the basic structure of web crawler[12], the following Fig. 2 starts from the URL list called seed, recognizes the hyperlink within web pages, and then continuously repeats the process of renewing the URL link. The web crawler is used as various terms like bots, spiders, or robots.

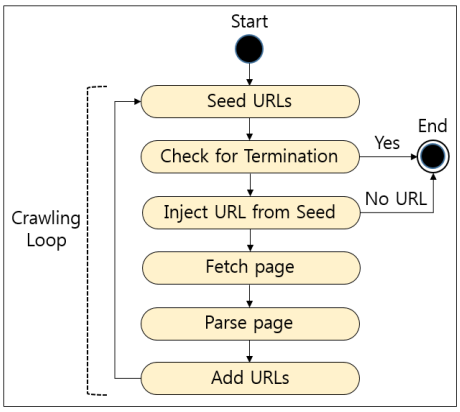

Fig. 2. Web Crawler Basic Structure

As a technique for analyzing big data by automatically collecting data of web pages, the data crawling method is various, and generally analyzes the information within web pages and utilizes the method of collecting numerous multimedia, texts, figures, and pictorial information included in them. Thus, the crawling system could be the system of collecting web pages by performing the works like protocol, encoding, DNS caching, and handling[13].

2.3 Logistic Regression

As a logistic regression model used for predicting the probability of events by matching data to the logistic curve[14], the logistic regression model is used in various fields such as genetics[15]. Also, the logistic regression is a type of regression analysis that could predict the categorical results based on a specific predictor variable[16]. Especially, the logistic regression model is very usefully used for classifying two types of classes. Supposing that there are samples(n=N) with two variables like , each sample could be viewed as binary class label of i=1, and the probability that could occur in the logistic regression of binary classification could be modeled as the following [Formula 1][17].

$$P(y=\pm 1 \mid x,\ w)=\frac{1}{1+\exp(-y(w^{T}x+b))} \quad (1)$$

Here, the intercept and level coefficient vector is the mediating variable to be assumed. Even though the logistic regression is based on a relatively simple theory, especially, as a model that could perform the probabilistic analysis on categorical data, this model could predict the matter of occurring a specific event by using an independent variable that has direct effects on a dependent variable.

3. Proposed Model

3.1 Model Configuration

The following Fig. 3 is the system structure of the proposed model. The AI recommendation system manages the analyzed data, and also performs the membership registration, up-loading notices, environment set-up, visual expression of strengths and weaknesses of products analyzed in system, and the logistic regression analysis prediction.

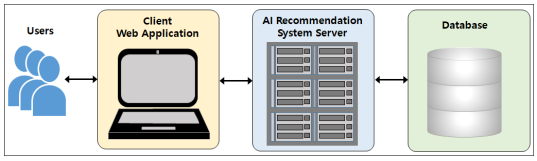

Fig. 3. Proposed Model System Structure

3.2 System Architecture

As an architecture of the proposed system, the following Fig. 4 is largely composed of three parts.

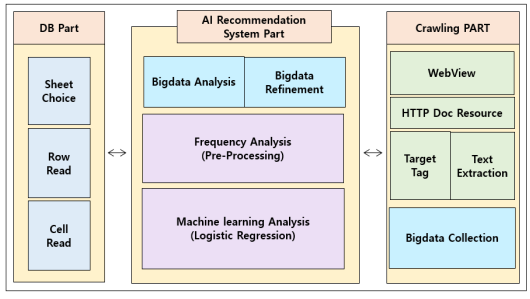

Fig. 4. System Architecture

First, as the database part, the detailed analysis data formed in the AI recommendation system part is saved, so that the analysis contents are visualized through web application. Second, as the AI recommendation system part, the unstructured data collected from the web crawling part is changed to structured data. It forms data that visualizes the strengths and weaknesses by analyzing the frequency through the natural language processing, and also predicts the satisfaction of target product through logistic regression. Third, a large-amount of data is collected by selecting a specific product for collecting big data through web crawling.

3.3 System Flowchart

The following Fig. 5 is the AI recommendation system process of the proposed model, which is largely composed of five steps.

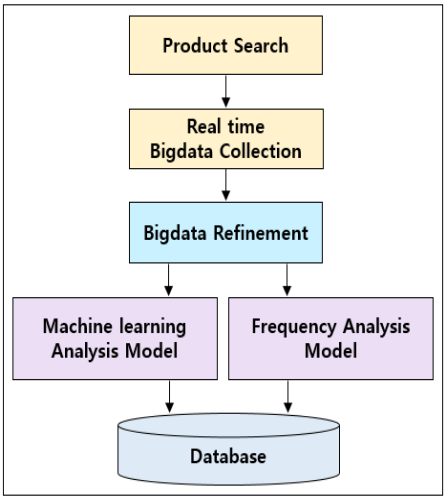

Fig. 5. System Flowchart

First, the site, target product, and period for collecting big data are selected. Second, the big data is collected in real time. Third, the unstructured data collected in real time is changed to the structured data for logistic regression and natural language processing. Fourth, the product satisfaction is predicted through the logistic regression analysis model. Fifth, the strengths and weaknesses are visualized by analyzing the frequency through the natural language processing process.

3.4 Text Mining

The following Fig. 6 is the text mining process, which could be largely divided into four steps. First, this data collection step performs the process of collecting the unstructured data stored in the massive digital form created in daily life through the web crawling technique, and also performs the process of cleansing the collected unstructured data into the structured data, so that the text mining technique could be utilized. Second, this text pre-processing step includes the process of tokenizing the texts and removing postpositions, special characters, and words unnecessary for the natural language processing. As a process of separating documents into sentences with punctuation marks, the tokenizing includes the process of removing punctuation marks, changing uppercases to lowercases, or tokenizing words and sentences. In case of Hangeul(Korean alphabet), the tokenizing considering the parts of speech is necessary. After that, it performs the process of dividing sentences into morpheme which is the smallest unit that has a meaning, and then attaching the information of part of speech, and the morpheme

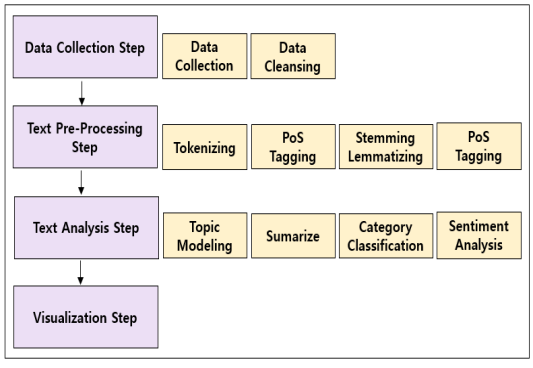

Fig. 6. Text Mining Process

analysis process of removing unnecessary parts of speech for analysis or filtering necessary parts of speech. It also performs the process of cognizing the entity name which is the process of distinguishing words that should be cognized as individual entity.

It uses Clunking that is a technique of binding information with significant words as one of the natural language processing techniques or utilizing the search engine index. It performs the process of restoring the original form that could extract the stem as the basic form of word, and also performs the process of stopword processing for unnecessary or disturbing words for analysis. Third, this text analysis step which detects the polarity within documents, paragraphs, and sentences, is a lot quoted as sentiment analysis data as the users express their sentiment in blogs, reviews, surveys, and social media. The sentiment analysis type uses the logistic regression as an emotion detection technique, and deep-learning algorithm, and there is a sentiment analysis method based on the characteristics within texts as an aspect-based sentiment analysis technique. In order to utilize the sentiment detection technique, this thesis changed it into structured data for creating data set for learning attached with sentiment label in the data cleansing process. Fourth, this visualization step performs the process of visualizing word cloud and logistic regression analysis prediction rate through the frequency analysis.

3.5 Algorithm

The following Fig. 7 is a part of logistic regression analysis learning algorithm of the proposed model, which is processed in the order of five steps as follows. First, after making a vocabulary dictionary of provided data and calculating the frequency of each word, the entity to be vectorized is formed. Second, the learning is performed by analyzing, digitizing, and vectorizing the number of words shown in each sentence. Third, the learning is performed. Fourth, the learning accuracy is measured. Fifth, the learnt model is saved.

```
def step2_learning(X_train, y_train, X_test, y_test):
    tfidf = TfidfVectorizer(lowercase=False, tokenizer=tokenizer)
    logisticRegression(C=10.0, penalty='l2', random_state=0)
    pipe = Pipeline([('vect', tfidf), ('clf', logistic)])
    pipe.fit(X_train, y_train)
    y_pred = pipe.predict(X_test)
    print(accuracy_score(y_test, y_pred))
    with open('pipe.dat', 'wb') as fp :
        pickle.dump(pipe, fp)
    print('save')
```

Fig. 7. Logistic Regression Algorithm

4. Test and Results

4.1 Proposed Model Main Screen

The following Fig. 8 is the main screen of the AI recommendation system management model. To use the system by connecting to the system of AI recommendation system model, the membership registration is essential because it is the system for product developers to analyze products. For the identification, the membership registration is progressed through the user authentication step, after joining as a member, the AI recommendation system management model could be used.

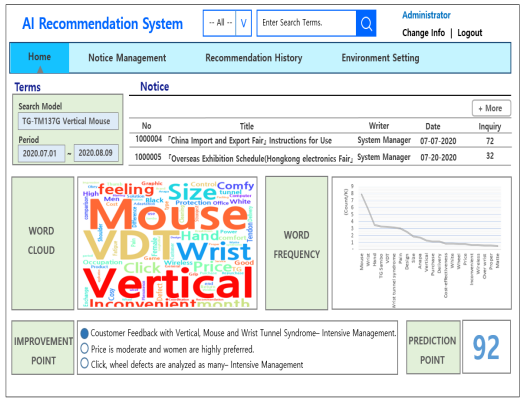

Fig. 8. Proposed Model Main Screen

4.2 Service Evaluation

Actually operating and using the recommendation model proposed by this thesis during the test period for 40 days, this study performed the user service evaluation of the proposed model by conducting a survey on the user satisfaction, system effectiveness, and system efficiency. The following Table 1 summarizes the contents of service evaluation. In order to evaluate the use of AI recommendation system, the items were classified into user satisfaction, system effectiveness, and system efficiency, and the detailed evaluation questions for each item were written. Through this process, a survey was conducted targeting total 50 product planners related to product sourcing. The following Table 2 is the survey results. In the results of the survey on the AI recommendation system, regarding the system quality, system effectiveness, system efficiency, and user satisfaction, the satisfaction or up was average 90.6%, and they were positive toward the proposed model system.

Table 1. General Matters of Service Evaluation [N=50]

Division	User Satisfaction	System Efficiency	System Effectiveness
Very Satisfied	37	35	35
Satisfied	8	11	10
Average	4	3	4
Dissatisfied	1	1	1
Very Dissatisfied	0	0	0

Table 2. Survey Results

Division	User Satisfaction
Research Method	Survey targeting the users of AI recommendation system mode
Test Period	July 1st 2020. ~ August 9th 2020.
Age of Survey Subjects	20s ~ 60s
Subjects	Product planners & people related to product sourcing

As the complementary matters, on top of general seasonal factor, there was an opinion about more elaborate demand prediction model for the active modification and complementation of data on climate changes in each season.

5. Conclusion

To conquer the rapidly-changing on/offline new product market, the planning and development of new products are playing huge roles in corporate performance and growth. One of the important things in the area of developing new products, is the joint participation of consumers. The consumers need to properly evaluate the satisfied needs of the existing products, and ideas of new services/products, which should be expanded to business model. This thesis conducted the word frequency analysis and the logistic regression analysis algorithm as a field of logistic regression to predict a product with the highest purchase possibility, which was used for the newest marketing technique for the improvement of problems with released products, the next best action or product recommendation, by analyzing the massive unstructured consumer data of portal sites and online markets, positiveness/negativeness, and predicting/analyzing the causes. In the results of logistic regression analysis of 11,200 data sets of a specific product collected from portal sites and online markets, the customer satisfaction was predicted as 92%. In the frequency analysis, there were defective issues of product, which should be urgently improved. It was analyzed as the contents that should be considered when releasing products in the future. This could have positive effects through corporate decision-making in case when releasing new products, and complementation through customer feedbacks after release. In the future direction of researches, there should be continuous researches on the realization of the proposed model and the realization of product development process

reflecting the technique with the strengthened accuracy of consumer sentiment analysis.

REFERENCES

[1] H, Evanschitzky., M., Eisend, R, Calantone, (2012). Success factors of product innovation: An updated meta-analysis. *Journal of product innovation management*, *29*, .21-37. DOI:10.1111/j.1540-5885.2012.00964.x

[2] K. B. Kahn, Barczak, G., Nicholas, J., Ledwith, A., & Perks, H., 2012. An examination of new product development best practice. J*ournal of product innovation management, 29(2),* 180-192. DOI:10.1111/j.1540-5885.2011.00888.x

[3] M. J. J. Lin, & C. H. Huang. (2013). The impact of customer participation on NPD performance: The mediating role of inter-organisation relationship. *Journal of Business & Industrial Marketing, 28(1),* 3-15.

[4] P. Trotter, & J. Vaughan. (2012). Innovation in UK companies. An evaluation of the implementation of best practice in front end innovation processes and methodologies. *International Journal of Innovation Science, 4(4),* 191-204. DOI:10.1260/1757-2223.4.4.191

[5] S. H. Kim, & K. H. Huarng. (2011). Winning strategies for innovation and high-technology products management. *Journal of Business Research, 64(11),* 1147-1150. DOI:10.1016/j.jbusres.2011.06.013

[6] B. Jeong, J. Yoon, & J. M. Lee. (2019). Social media mining for product planning A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management, 48,* 280-290. DOI:10.1016/j.ijinfomgt.2017.09.009

[7] D. Silverstein, P. Samuel, & N. DeCarlo. (2013). The innovator's toolkit: 50+ *techniques for predictable and sustainable organic growth.* John Wiley & Sons.

[8] W. Medhat, A. Hassan, & H. Korashy, (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal, 5(4),* 1093-1113. DOI:10.1016/j.asej.2014.04.011

[9] B. Pang, & L. Lee. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058.* DOI: 10.3115/1218955.1218990

[10] S. M. Kim, & E. Hovy. (2004). Determining the sentiment of opinions. *In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics,* 1367-1373.

[11] A. Agarwal, F. Biadsy, & K. Mckeown. (2009). Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. *In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 24-32.

[12] G. Pant. P. Srinivasan. & F. Menczer. (2004). Crawling the web. *In Web dynamics, Springer, Berlin, Heidelberg*, 153-177.

[13] V. Shkapenyuk, & T. Suel. (2002). Design and implementation of a high-performance distributed web crawler. *In Proceedings 18th International Conference on Data Engineering, IEEE*, 357-368. DOI:10.1109/ICDE.2002.994750

[14] F. E. Harrell. (2015). Ordinal logistic regression. *Regression modeling strategies, Springer, Cham,* 311-325.

[15] C. M. Lewis, & J. Knight (2012). Introduction to genetic association studies. *Cold Spring Harbor Protocols, 2012(3)*, pdb-top068163.

[16] S. Agarwal, P. Kachroo, & E. Regentova. (2016). A hybrid model using logistic regression and wavelet transformation to detect traffic incidents. *Iatss Research, 40(1),* 56-63. DOI:10.1016/j.iatssr.2016.06.001

[17] W. Chen, et al. (2018). GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. *Science of the Total Environment, 634,* 853-867. DOI:10.1016/j.scitotenv.2018.04.055

안 영 휘(Hwi-Yeong Ahn) [정회원]

· 2005년 2월 : 국립한밭대학교 컴퓨터공학과(공학사)
· 2007년 2월 : 국립한밭대학교 컴퓨터공학과(공학석사)
· 2019년 3월 ~ 현재 : 국립공주대학교 컴퓨터공학과(박사수료)
· 관심분야 : 인공지능, 데이터 사이언스, IT 컨버전스, 컴퓨터비전, 상품기획
· E-Mail : yhahn02@naver.com

박 구 락(Koo-Rack Park) [정회원]

· 1986년 2월 : 중앙대학교 전기공학과(공학사)
· 1988년 2월 : 숭실대학교 전자계산학과(공학석사)
· 2000년 2월 : 경기대학교 전자계산학과(이학박사)
· 1991년 ~ 현재 : 공주대학교 컴퓨터공학부 교수
· 관심분야 : IT 컨버전스, 정보통신, 머신러닝, 전자상거래
· E-Mail : ecgrpark@kongju.ac.kr

김 동 현(Dong-Hyun Kim) [정회원]

· 1986년 2월 : 중앙대학교 전기공학과 (공학사)
· 2005년 2월 : 공주대학교 컴퓨터멀티미디어공학과(공학석사)
· 2010년 2월 : 공주대학교 컴퓨터공학과(공학박사)
· 2021년 2월 ~ 현재 : 나사렛대학교 IT 인공지능학부 교수
· 관심분야 : 인공지능, 영상처리, 지식관리, 시뮬레이션
· E-Mail : dhkim@kornu.ac.kr

김 도 연(Do-Yeon Kim) [정회원]

· 2014년 2월 : 국립충남대학교 항공우주공학과(공학사)
· 2018년 8월 : 국립공주대학교 멀티미디어공학과(공학석사)
· 2019년 3월 ~ 현재 : 국립공주대학교 컴퓨터공학과 (박사수료)
· 관심분야 : IT 컨버전스, 자동제어, 빅데이터, 인공지능
· E-Mail : kdy4922@naver.com