❒   5454

Product recommendation system based user purchase criteria and product reviews

Jinyoung Kim, Doyeun Hwang, Hoekyung Jung
Department of Computer Engineering, Pai Chai University, South Korea

Article Info	ABSTRACT
	In this paper, we propose a system that provides customized product recommendation information after crawling product review data of internet shopping mall with unstructured data, morphological analysis using Python. User searches for a proudct to be purchased and select the most important purchase criteria when purchasing the product. User searches for a proudct to be purchased and select the most important purchase criteria when purchasing the product. And extracts and analyzes only the review including the purchase criterion selected by the user among the product reviews left by other users. The positive and negative evaluations contained in the extracted product review data are quantified and using the average value, we extract the top 10 products with good product evaluation, sort and recommend to users. And provides user-customized information that reflects the user's preference by arranging and providing a center around the criteria that the user occupies the largest portion of the product purchase. This allows users to reduce the time it takes to purchase a product and make more efficient purchasing decisions.

Corresponding Author:

Hoekyung Jung,
Department of Computer Engineering,
Pai Chai University,
155-40 Baejae-ro, Seogu, DaeJeon, South Korea.
Email: hkjung@pcu.ac.kr

1. INTRODUCTION

As the Internet and the mobile environment develops, not only the amount of documents but also the types of documents and the purpose of the documents are becoming diverse [1-4]. Among them, reviews of products are written by several consumers, and the amount of data is rapidly increasing, and it is shared according to the characteristics of open web [5, 6]. These product reviews contain detailed and reliable information about the user's preferences about the product. Therefore, this method can be applied to the recommendation system [7-10], and various methods of analyzing the vast amount of suggested information in the product review have been proposed [11-13].

However, most of the filtering techniques used in the recommendation system uses the information of the user to provide the appropriate information, but there is a problem of the cold start which can not provide proper information when a new user or a new product is released [14, 15]. And The existing product recommendation system is based on information that is not objective or has low satisfaction because it does not take into consideration user 's preference [16]. Therefore, if the user considers the criteria that are important when purchasing the product, the satisfaction level of the recommended product can be increased.

In order to solve this problem in this paper, the user directly selects the criteria of 'durability', 'design', and 'cost-effectiveness'. The top 10 products are sorted and recommended to the user after extracting the products having excellent evaluation on the purchase criteria using the purchase criteria selected by the user and the product reviews created by the users who has previously purchased the products.

Using this analysis result in purchasing decision making, it can save time for purchasing product and help efficient decision making.

2.   SYSTEM DESIGN

First, a user searches for a product to be purchased, and selects one of the purchase criteria that is considered as a top priority when purchasing the product. The product retrieved by the user and the selected purchase criterion are used as input data for further analysis. This system collects a list of the products searched by the user and a review of the product. We use the Python Webdriver package to crawl the product's URL, product name, and product reviews.

Then, the collected input data is stored in the database. In case of 'G' company, which is the target of data collection, there is 'premium product review' where users can write specific product reviews. We performed filtering based on the number of product reviews corresponding to 20 of these premium product reviews. 20 premium product reviews are the number of 4 page reviews. Products with fewer than 20 product reviews were judged to be insufficient to analyze and were excluded from the analysis.

When extraction is completed, we used the konlpy and kkma packages of Python to perform morpheme analysis. Thereafter, a product list including purchase criteria selected by the user is generated. The list is divided for each product to calculate the average value of positive and negative scores for the product reviews. The results are sorted in ascending order according to the result value, and the list of the top 10 products is provided to the user together with the score obtained as a result. Figure 1 shows the structure of the system and Figure 2 shows the system flow chart.
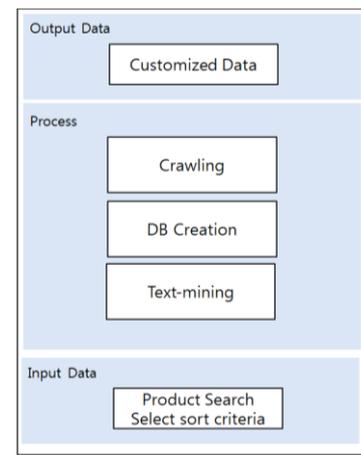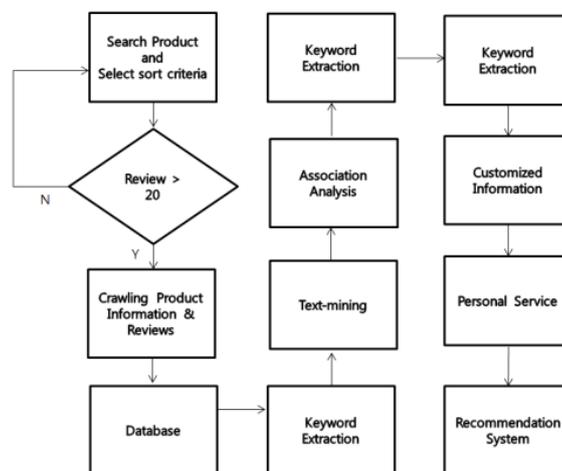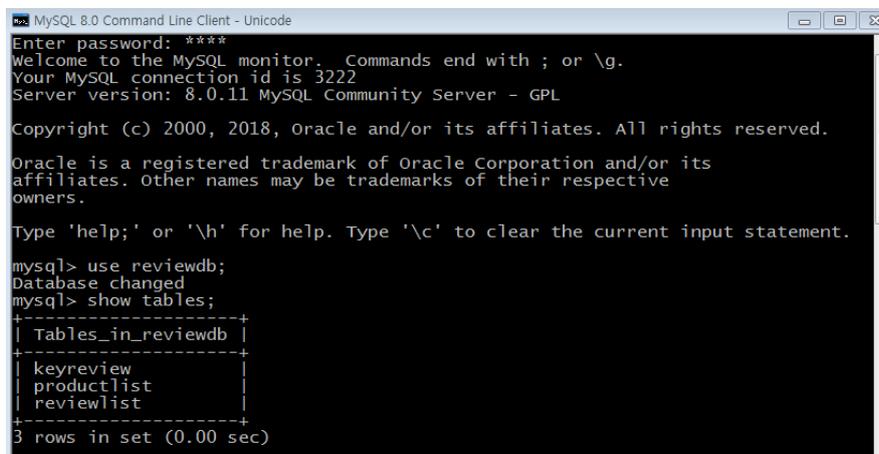

Figure 1. System architecture

Figure 2. System flowchart

The database of this paper is designed using MySQL. Database created three tables in one database called 'Crwal'. One of the tables is 'productlist', and stores product information such as an ID and a product name given for each product, and a URL of the corresponding product. The ID of the product may be a unique number assigned to each product in the Internet shopping mall. This is useful for a product with the same brand name through ID when searching for a product, and it plays a role of matching reviews and products. And it is possible to connect directly to the product purchase site through URL.

Another table has the name of 'reviewlist' and loads the product ID and product reviews. Product review analysis should be preceded by finding the score of the sentence to which each feature belongs. To do this, we need to separate sentences. It is common for each review to consist of one less sentence, more than 10 sentences. Therefore, when crawling product reviews, by storing sentences based on a period, they are useful for later morphological analysis.

The final 'keyreview' table extracts and stores product reviews that include the purchase criteria selected by the user and the word related to the criteria. And then it removes the meaningless words such as 'ㅋ ㅋㅋㅋ' or special characters. Figure 3 shows the designed DB and table.

Figure 3. Creation database and tables

Figure 4 shows that the privileges are set for each table. It is a function that is needed when performing query using data collected and analyzed by Python. Therefore, each table is granted the right to select, insert, update, delete, etc.
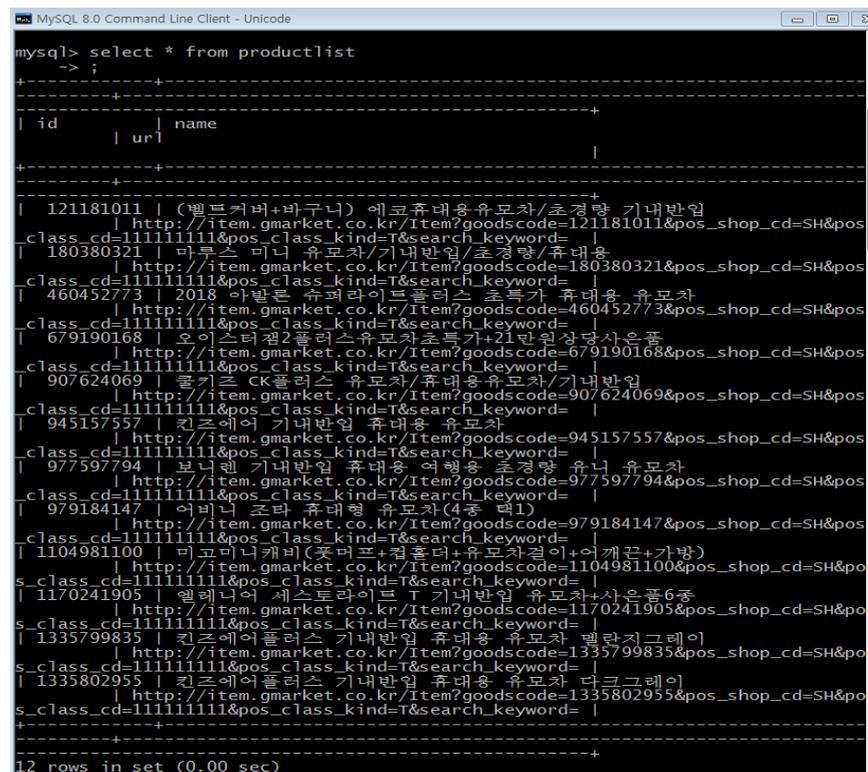

Figure 4. Grant privileges on tables

3. SYSTEM IMPLEMENTATION

We used CPU-Intel i5-4690 and RAM-8G based on Windows operating system. The language used is Python and MySQL is used to build the database, and a case where the user selects 'cost-effectiveness' as a purchase criterion is implemented as an example.

The data collection site was selected as 'G Company' in Korea. It collects reviews made by the user about product information and products only for "Baby Carriage" of childcare item among the categories of the largest shopping mall in Korea. Since the quality of the product reviews influences this experiment, we chose the category as a childcare item that is relatively straightforward and has many written reviews. It is collected using BeautifulSoup, a library of Python that allows you to crawl product names, product URL, and product reviews of 'stroller' items.

However, when reviews of more than a certain number were writtened on the online shopping mall system, all of them can be seen by pressing the "More" button. Likewise, when you crawl, you must be able to see a list of all reviews if an action is taken on the web browser to page through. So, we use Selenium, a web page testing tool that can be performed automatically when we crawl these actions. The Selenium package functions to launch the web browser through a product URL in a Python environment. Crawl the product name, product classification, review, etc. using the HTML tag name.

The product reviews that were crawled previously include 'caustic rain' and 'price', and only the number of reviews exceeding 20 is extracted and stored. As a result, 12 items were extracted. Figure 5 shows the list of 12 items stored in DB. The purchase criteria 'cost-effectiveness' and the related keyword 'price' extract only the name of the product including the review sentence.

Figure 5. List of extracted product

For product reviews crawled, install 'KoNLP', a package for processing Hangul characters of Python programs. The morphological analysis is based on the KKMA dictionary and is carried out to extract only nouns and adjectives from each product review sentence. The part-of-speech which can judge the characteristic of the product is the noun, and the part-of-speech which can express the sensitivity of the user to the characteristic is the adjective. In this process, sentences that do not contain adjectives that can judge positive or negative of the sentences of the review are judged as noise and excluded from the analysis.

Figure 6 shows an example of a review of morpheme analysis by extracting only the reviews including the keyword 'price' associated with the cost-effectiveness by selecting the sorting criterion 'cost-effectiveness'.

Figure 6. Example of morphemic analysis using KKMA dictionary

In order to recommend a product, It is necessary to judge which evaluation is more positive or negative among the product reviews extracted previously. Therefore, in the case of a product review, a score of "+1" is given if there is a positive word to evaluate it in a sentence containing the keyword or the creteria, or "-1" is given if there is a negative word. Then the average of the sum of positive and negative scores is divided by the total number of product reviews (C).

In this case, if the set of collected reviews $R = \{r_1, r_2, ..., r_n\}$ and the set of emotional words $S_{r_n} = \{s_1, s_2, ..., s_n\}$ resent in any review, The emotion score $SentenceGrade(r_n)$ of any review, which is the sum of the positive score (P) and the negative score (N) divided by the total number of reviews, is expressed as the following (1).

$$SentenceGrade(r_n) = \frac{P(s_n) + (-N(s_n))}{p_n} \qquad (1)$$

However, since neutral opinions were difficult to distinguish, they were judged to be noise and excluded. And overall product reviews are positive, but included in the rating exclusion list, which does not have a positive srore on the sorting keyword 'price' or 'cost-effectiveness'. Table 1 shows the positive and negative scores of each of the twelve product reviews extracted, and is a table summarizing the averages. Table 2 is a table ranked according to the average of positive and negative scores of the product reviews extracted. It is a list of products recommended to the user.

Table 1. Result of calculated means

Product ID	Count of Review	Result Score(P+N)	Average((P+N)/C)
1335802955	8	5	0.625
1211801011	6	6	1.0
1335799835	4	3	0.75
1803280321	16	15	0.9375
679190168	3	3	1.0
945157557	8	7	0.875
907624069	25	24	0.96
979184147	60	54	0.9
977597794	10	10	1.0
460452773	18	13	0.72
1170241905	9	8	0.89
1104981100	4	3	0.75

Table 2. Rank table with scores

Ranking	Product ID	Score
1	97797794	1.0
2	121181011	1.0
3	679190168	1.0
4	907624069	0.96
5	180380321	0.937
6	979184147	0.9
7	1170241905	0.89
8	945157557	0.875
9	1104981100	0.75
10	460452773	0.72

4.   REVIEW

Table 3 shows the contents of the proposed system. System A calculated positive words, negative words, key words, and so on from the data extracted from the advertisement through web parsing in SNS. And designed a system that displays the analysis results on a web server using a distributed system [17]. Because collect data from the SNS, it can collect and analyze data in real time, the amount of accumulated data increases exponentially, so there is a concern that accuracy and reliability may be reduced. Incorrect information may be included in the calculation as a weighting factor, which may or may not be accurate.

System B analyzes the product reviews containing the consumer's ratings using natural language processing analysis and automatic classification system to automatically identify the consumer's preferences. The system Identify the disadvantages of the product that the consumer thinks and designed a system that assigns a quantitative comprehensive score that reflects it to product reviews [18]. In the system, classify product reviews into five classification codes such as very positive, positive, tentative, negative, and very negative. And then it is necessary to analyze it to clarify whether the emphasized word emphasizes the meaning of positive or negative.

System C complements the problem that it takes much time to extract sentiment words in Korean language research and suggests a technique that can be directly applied to Korean [19, 20]. It is a k-structure, proposed a method of extracting sentiment words and attribute words in a simple Korean product review [21]. But it is insufficient to reflect the attribute according to the characteristic of the products because the emotional word is extracted by fixing the eight attribute words. And there are limitations in analyzing longer sentences because they are limited to relatively simple sentences with a maximum pattern length of 3.

System D compensate the defect that limited information that can be given by raw data and suggests a system that provides processed information through emotion analysis of review sentences as well as basic information.

Table 3. Table of comparsion existing systems

Systems	Data to analyze	Keyword
A	SNS review data about cosmetics	Positive score, Negative score, Analogize word
B	Product review data	5 Classification codes
C	Korean product review	Property word, Emotional word
D	Domestic internet bookstore review data	Modifier word, Emphasis word, Keyword
E	Review data from Smartphone app marketplace	User-to-user review similarity, Similarity score between users
Proposed System	Review data for shopping mall G baby carriage	User purchase criteria

In this way, the problems of existing systems that do not reflect the classification and attributes of goods are improved. In this system, proposed a system that provides e-commerce product analysis and evaluation by review emotion analysis that reflects classification and attributes of products. Collect product review data of domestic bookstore shopping mall site and Implement a system that visually provides the results through review emotion analysis. It separates the review data into key words, modifiers, and emphasis words by a morpheme analyzer and the evaluation of the product is calculated as the emotion analysis algorithm [22]. It provides a visual representation of the results of the analysis, which helps the user make a buying decision but it does not reflect user's preference is insufficient in terms of role as recommendation system.

System E proposes a new recommendation algorithm that can improve the performance of the collaborative filtering recommendation technique. It refers to the reviews written by the user without considering only the score. In this system, the similarity degree of the user review is calculated by the frequency of the index words commonly extracted from the review data or the weight sum of the corresponding index words [23]. This system collects and analyzes review data from the smartphone app marketplace and recommends it to users. This is a differentiating part because it takes into consideration the similarity of user reviews, which is qualitative information. But the prediction accuracy of the recommendation algorithm considering user reviews is lower than the comparison algorithm.

Most consumers pay a certain amount of money and purchase goods by necessity. But the criteria for selecting a product are different for each individual. The criteria for purchasing goods is information that only the purchaser knows so, in order to reflect this, in this system, users were selected together with product search. Calculates the similarity with other users or the frequency or weight of a specific word and it reflects the purchasing criteria selected by the user unlike the existed recommended system. The most differentiating part is that generated customized information for each new individual.

5.   CONCLUSIONS

Most consumers consider product reviews and it reflect their purchasing decisions [24, 25]. But existing online shopping mall sites have a problem that it takes a lot of time for a consumer to search for a product and to select a desired product among many kinds of products. The biggest reason that the greatest amount of time and effort for purchasing is the amount of unnecessary information is huge, and the criteria for selecting a product are different for each individual. But despite having different purchase criteria, the same product list is provided, which leads to inefficient purchasing decisions. Therefore, in this research the user can select the purchase criterion along with the product search and reflect it and by reflecting this, we complemented the inefficient purchase decision process in which product information is provided randomly.

The ultimate goal of a product recommendation system is to help users make purchasing decisions by recommending products that are really needed by users to the right situation. It is also important for users to know what they need in order to recommend products to users but understanding the hidden meaning of the reason for purchasing a product can enhance the user's satisfaction. In order to increase the satisfaction of the users, a method of recommending products based on the purchasing criteria selected by each user is needed. Therefore, in this system rearranges the product list that reflects the products the user searched for and the selected purchase criteria. This not only reduces the time it takes to purchase a product, but also filters information reflecting its own purchase criteria and provides information including keywords selected by the user.in order of product rating, and provides it to the user.

Also, since the filtered list is rearranged in the order of good evaluation, there is no need to read a large amount of product list and product review. It can be expected that users will be able to receive product information quickly according to a wanted purchase criterion, and increasing the satisfaction of the user. However, there is also a case where the product review data is created with malicious intention rather than a product review made by purchasing and directly using the product. This needs to be filtered because it interferes with the user's decision to buy the product. Future research will be able to show how to filter false review data.

REFERENCES

[1]   J. Y. Chang, "A Sentiment Analysis Algorithm for Automatic Product Reviews Classfication in On-Line Shopping Mall," *The Journal of Society for e-Business Studies*, vol. 14, pp. 19-33, 2009.

[2]   A. Esuli and F. Sebastiani, *"Determining Term Subjectivity and Term Orientation for Opinion Mining,"* Proceedings of 11th conference of the European chapter of the Association for Computational Linguistics : EACL, pp. 193-200, 2006.

[3]   E. J. Yu, et al., "Predicting the Direction of the stock index by using a domain-specific sentiment dictionary," *Journal of intelligent information systems*, vol. 19, pp. 95-100, 2013.

[4]   X. Ding, et al., *"A holistic lexicon-based approach to opinion mining,"* Proceedings of the international conference on Web search and web data mining, pp. 231-240, 2008.

[5]   S. Morinaga, et al., *"Mining Product Reputations on the Web*," Proceedings of the 8th SIGKDD International Confer- ence on Knowledge Discovery and Data Mining, pp. 341-349, 2002.

[6]   H. G. Lee and H. Kwak, "Investigation of factors affecting the effects of online consumer reviews," *Journal of national information society agency,* vol. 20, pp. 3-17, 2013.

[7]   Chang J. Y., "A Sentiment Analysis Algorithm for Automatic Product Reviews Classfication in On-Line Shopping Mall," *The Journal of Society for e-Business Studies*, vol. 14, pp. 19-33, 2009.

[8]   S. M. Alex, et al., "Development of Deep Learning Models for Multi-class Sentiment Analysis," *Journal of Information Technology Services*, vol. 16, pp. 149-160, 2017.

[9]   S. M. Ahn, et al., "Understanding the Performance of Collaborative Filtering Recommendation through Social Network Analysis," *Journal of Society for e-Business Studies*, vol. 17, pp. 129-147, 2012.

[10]   Z. Zhang, et al., "A hybrid fuzzy-based personalized recommender system for telecom products/services," *Information Science,* vol. 235, pp. 117-129, 2013.

[11]   K. R. Jang, et al., "Extracting lmplicit customer viewpoints from product review text," *Journal of Korea information science society: software and applications*, vol. 41, pp. 376-386, 2014.

[12]   J. Y. Kim and D. S. Kim, "A study on the method for extracting the purpose-specific customized information from online product reviews based on text mining," *Journal of society for e-business studies*, vol. 21, pp. 151-161, 2016.

[13]   K. Cao, J. Lee, and H. Jung, "Keyword Analysis Based Document Compression System," *Journal of information and Communication Convergence Engineering,* vol. 16, no. 1, pp. 48-51, 2018.

[14]   I. Y. Jeong, et al., "A Study on Movies Recommendation System of Hybrid Filtering-Based," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 19, pp. 113-118, 2015.

[15]   J. Y. Hyun, et al., "How to improve the accuracy of recommendation systems: Combining ratings and review texts sentiment scores," *Journal of Intelligence and Information Systems*, vol. 25, pp.219-239, 2019.

[16]   B. Jeon and H. Ahn, "A Collaborative Filtering System Combined with Users Review Mining: Application to the Recommendation of Smartphone Apps," *Journal of Intelligence and Information System*, vol. 21, pp. 1-18, 2015.

[17]   A. R. Lee, et al., "A Design of a TV Advertisement Effectiveness Analysis System Using SNS Big-data," *Journal of KIISE Transactions on Computing Practices*, vol. 7, pp. 579-586, 2015.

[18]   G. Y. Kim, et al., "Customer Preference Identification System using Natural Language Processing-based Analysis and Automatic Classification of Product Reviews," *Journal of Korea Society for Information Management,* vol. 5, pp. 65-70, 2009.

[19]   An J. K. and H. W. Kim, "Building a korean sentiment Lexicon using collective intelligence," *Journal of Intelligent Information Systems*, vol. 21, pp. 46-67, 2015.

[20]   M. Hu and B. Liu, "Mining and summarizing customer reviews," *Proceedings of the 10th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pp. 168-177, 2004.

[21]   H. H. Kang, et al., "Automatic Extraction of Opinion Words from Korean Product Reviews Using the k-Structure," *Journal of KISS: Software and Applications*, vol. 10, pp. 470-479, 2010.

[22]   J. E. Choi, et al., "System Design for Analysis and Evaluation of E-commerce Products Using Review Sentiment Word Analysis," *Journal of KIISE Transactions on Computing Practices*, vol. 22, pp. 209-217, 2016.

[23]   B. K. Jeon and H. C. Ahn, "A Collaborative Filtering System Combined with User's Review Mining," *Journal of Intelligence and Information Systems*, vol. 19, pp. 25-44, 2018.

[24]   A. Esuli and F. Sebastiani, "Determining Term Subjectivity and Term Orientation for Opinion Mining," *Proceedings of 11th conference of the European chapter of the Association for Computational Linguistics: EACL*, pp. 193-200, 2006.

[25]   Z. Huang, et al., "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering," *ACM Trans. Information Systems*, vol. 22, pp. 116-142, 2004.

BIOGRAPHIES OF AUTHORS

Jinyoung Kim received the B.S. degree in 2001 and M.S. degree in 2011 from the Department of Information Technology of Chonubk National University, Korea. Since 2001, he has worked for K-water as a manager especially in Water Treatment Plants. He is currently in a Doctorate course in Department of Computer Engineering of Paichai University. His current research interests include 5G, Cloud Computing, Machine Learning, Bigdata and Artificial Intelligence.

Doyeun Hwang received her B.S., degrees from the Department of Business Administrarion of Paichai University, Korea, in 2017. She has been studying in the Department of Computer Engineering at Paichai University. Her current research interests include text-mining, bigdata, opinion-mining.

Hoekyung Jung received the M.S. degree in 1987 and Ph. D. degree in 1993 from the Department of Computer Engineering of Kwangwoon University, Korea. From 1994 to 1995, he worked for ETRI as a researcher. Since 1994, he has worked in the Department of Computer Engineering at Paichai University, where he now works as a professor. His current research interests include multimedia document architecture modeling, information processing, information retrieval, machine learning, bigdata, and IoT.