ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org

http://dx.doi.org/10.13088/jiis.2020.26.1.151

A Hybrid Collaborative Filtering-based Product Recommender System using Search Keywords

Yunju Lee
Graduate School of Business IT,
Kookmin University
(*mmlas0ui@kookmin.ac.kr*)

Haram Won
Graduate School of Business IT,
Kookmin University
(*haramy44@kookmin.ac.kr*)

Jaeseung Shim
Graduate School of Business IT,
Kookmin University
(*simong_kwanu@naver.com*)

Hyunchul Ahn
Graduate School of Business IT,
Kookmin University
(*hcahn@kookmin.ac.kr*)

A recommender system is a system that recommends products or services that best meet the preferences of each customer using statistical or machine learning techniques. Collaborative filtering (CF) is the most commonly used algorithm for implementing recommender systems. However, in most cases, it only uses purchase history or customer ratings, even though customers provide numerous other data that are available. E-commerce customers frequently use a search function to find the products in which they are interested among the vast array of products offered. Such search keyword data may be a very useful information source for modeling customer preferences. However, it is rarely used as a source of information for recommendation systems. In this paper, we propose a novel hybrid CF model based on the Doc2Vec algorithm using search keywords and purchase history data of online shopping mall customers. To validate the applicability of the proposed model, we empirically tested its performance using real-world online shopping mall data from Korea. As the number of recommended products increases, the recommendation performance of the proposed CF (or, hybrid CF based on the customer's search keywords) is improved. On the other hand, the performance of a conventional CF gradually decreased as the number of recommended products increased. As a result, we found that using search keyword data effectively represents customer preferences and might contribute to an improvement in conventional CF recommender systems.

Key Words : Recommender System, Hybrid Collaborative Filtering, Doc2Vec, Word Embedding, Search Keyword

Received : February 3, 2020 Revised : March 21, 2020 Accepted : March 24, 2020
Publication Type : Concise Paper Corresponding Author : Hyunchul Ahn

1. Introduction

Recommender systems, which are currently widely used by major online services such as Amazon, Netflix, and YouTube, analyze customer purchase history or ratings to recommend the products and services that an individual customer might prefer or need (Park *et al.*, 2009; Takács *et al.*, 2009). Especially for online businesses with many customers and a large amount of product data, a sophisticated recommender system that accurately predicts customer preferences can provide businesses with a competitive advantage since it might lead to higher customer satisfaction and an increase in sales (Kim and Kim, 2014).

In previous research on recommender systems, products were recommended to customers based on their purchase history or ratings. However, in a highly competitive online market, there are not many cases in which customers make re-purchases at shopping malls (Kim *et al.*, 2014). In addition, not all customers rate every product that they purchase. Therefore, recommendations that consider only purchase history or ratings data may be inaccurate. With this background, it is necessary to examine recommender systems that reflect customer preferences by observing new customer information, rather than simply tracking purchase history or ratings.

In online shopping malls, various information related to customer preferences can be found through an analysis of customer behavior information. In recommender system research, the most frequently used customer information, aside from their purchase history and ratings information, was customer reviews (Cho *et al.*, 2015; Choi and Ahn, 2017; Lee *et al.*, 2019). Customer reviews are information that can be easily obtained from online shopping malls, but they present some pre-processing difficulty.

On the other hand, customer search keyword data can be obtained from online shopping malls as easily as customer reviews and the pre-processing technique is not difficult because the majority of searched keywords are in noun form. Also, search keywords reflect customer preferences in that they are obtained from search behavior in order to acquire information about products that are being considered for purchase *before* the customer makes a purchase (Lee *et al.*, 2016). However, there are not many studies on recommender systems that utilize search keywords, so conducting research on this topic is necessary at this time.

In this paper, we propose a hybrid collaborative filtering (CF) methodology using the search keyword data of online shopping mall customers. We determined that if the keywords searched by various customers are similar, they represent different customers that have similar preferences. We intend to derive the similarities between customers using Doc2Vec and combine them with CF. In addition, the proposed model is expected to solve the cold start problem, which is a limitation of the conventional CF methodology, and to provide more accurate recommendation performance than the conventional method in which recommendations are based only on purchase history or ratings.

2. Theoretical Background

2.1 Recommender System and CF

The approach to recommendation systems is divided into content-based recommendation and CF (Balabanovic and Shoham, 1997; Billsus and

Pazzani, 1998). Content-based recommendation systems analyze the properties of an item and recommend a similar item based on the user's preferences. On the other hand, CF is a method of recommending to customers the products also preferred by other customers (Breese *et al.*, 1998; Kim & Kim, 2014; Park *et al.*, 2009). CF is considered to be a high-performance recommendation system in the industry and by academia; many researchers, such as Koo *et al.* (2017), have used customer purchase history data and ratings data as preference data (Ahn *et al.*, 2004; Kim, 2008; Ku *et al.*, 2017).

CF is, for the most part, classified into two groups: item-based CF and user-based CF. Item-based CF is a method that analyzes the similarities between items and recommends them. Item-based CF, therefore, is able to make continuous recommendations according to product purchase history. However, the similarity between users is not considered, so the recommendation quality might be lower for users who have different preferences. However, if the user's historical data about the product is insufficient, there might be a cold start, which refers to an initial lack of information. Since the total number of products is much larger than the number of products that a customer can possibly purchase, a sparsity problem arises in which the customer-product matrix becomes very sparse.

To solve these problems, many researchers have proposed a hybrid CF method that complements the advantages and disadvantages of each CF (Choi *et al.* 2019; Choi and Ahn, 2017; Shin *et al.* 2018). Recently, research has been conducted in Korea that proposes a recommender system using data other than customer purchase history or ratings.

Choi and Ahn (2017) employed user reviews and exploited Onion Mining to extract user preferences for specific products. Based on these results, a user's overall rating for each product was calculated. In addition, the results of sentiment analysis of online user reviews were applied to the recommender system. In order to deal with a lack of data, Choi *et al.* (2019) analyzed ratings data as well as behavioral information such as clicks, wish lists, and customer reviews to calculate user preferences. However, previous research has not been verified by empirical analysis and most studies obtained customer preferences using customer reviews.

In an overseas study, Stiebellehner *et al.* (2017) proposed a hybrid filtering model for app users based on document embedding. They collected app history and app description data from Apple's online app store and Google's Play store and analyzed these unstructured data using the Doc2Vec technique. They arrived at the conclusion that the quality of recommendations was improved when combined with CF.

With this background, this study proposes a new recommender system by deriving new customer preferences based on customer search keyword data—which have not been utilized often in previous research—and analyzing them using a hybrid CF method combined with Doc2Vec.

2.2 Doc2vec

Doc2Vec is an extended method of the Word2Vec algorithm, which finds associations between words in sentences and converts them into vectors. Doc2Vec makes a comparison between documents using a simple artificial neural network to find similar documents and then locates them closely in multi-dimensional spaces (Mikolov *et al.*, 2013). That is, each document is represented as a vector and each vector is trained to predict words in documents or sentences. There are two approaches to Doc2Vec: distributed memory (DM) and distributed bag of words (DBOW) (Le and Mikolov, 2014). First, the DM method, which is similar to the Word2Vec continuous bag of words method, combines the document identification (ID) vector and the word vectors of the document to predict other words in the document. In the DM method, a word vector inherits the meaning of the word and has the same meaning in all documents. The DBOW method predicts a document from the words contained in the document and ignores the sequence of words in the document. This method is similar to the Word2Vec Skip-Gram method (Le and Mikolov, 2014).

In Korea, the number of studies using Doc2Vec has not been large and the majority of the research has been published mainly on the topic of document embedding. Jung *et al.* (2019) embedded documents using Doc2Vec and proposed a document classifier using machine learning technology. Shim *et al.* (2019) proposed a method to classify each fake news document as ‘true’ or ‘false’ using Doc2Vec. Park and Kim (2019) recommended a new multi-vector document embedding method to solve the limitations of Doc2Vec’s document embedding and Ki *et al.* (2018) classified customers using Doc2Vec’s embedding characteristics. As such, Doc2Vec research has been used for document embedding and classification in Korea. However, there is no research that uses Doc2Vec in a recommender system.

However, it is possible to find overseas research that uses Doc2Vec in recommender systems. Phi *et al.* (2016) proposed a CF model based on Word2Vec and Doc2Vec and Stiebellehner *et al.* (2017) proposed two hybrid CF models based on Doc2Vec. Liu and Wu (2019) extracted feature vectors using sentences from movie synopses with Doc2Vec to derive similarities and then combined them into a recommender system. Karvelis *et al.* (2018) proposed a topic recommender system model that integrates Doc2Vec and recommends the document topic by extracting the topic based on document information. Karvelis *et al.* (2018) demonstrated that using the Doc2Vec method through the proposed model is better than the bag of words method, which is one of the natural language processing methods. Zhang *et al.* (2018) used Doc2Vec to recommend the web service most similar to the customer’s needs. In addition, Nandi *et al.* (2018) proposed a content-based Bangla news recommender system using paragraph vectors, also known as Doc2Vec, and found that the Doc2Vec technique enabled language-independent learning and adaptation capability with a large corpus. As such, the performance of recommender systems using the Doc2Vec technique can be observed through these overseas research studies.

2.3 Search Keyword Data

This research focuses on customer search keyword data from online shopping malls as unstructured data to be used in the model. According to Kang and Choi (2019), customer search behavior is based on the action of collecting information on products of interest, which may indicate customer preferences. In addition, in this research, customer search behavior can influence customer buying behavior.

Therefore, customer search behavior is very important information for online shopping malls. Online shopping malls focus on customer search behaviors and use a number of management strategies, such as using keyword marketing, to direct customers to their shopping mall sites or provide relevant search terms in response to their search behaviors.

Several studies have been conducted focusing on customer search behavior. A study focused on keyword marketing and suggested a method for automatically extracting search keywords that included customers' shopping intentions (Kim *et al.*, 2014); another study examined customer search history and found that more relevant recommendations were possible when search results were provided based on the customer's search history (Adomavicius and Tuzhilin, 2011). As such, customer search behavior can influence their buying behavior. Therefore, such data can be used to recommend products to a similar customer based on the purchase history of another customer whose search keywords are similar.

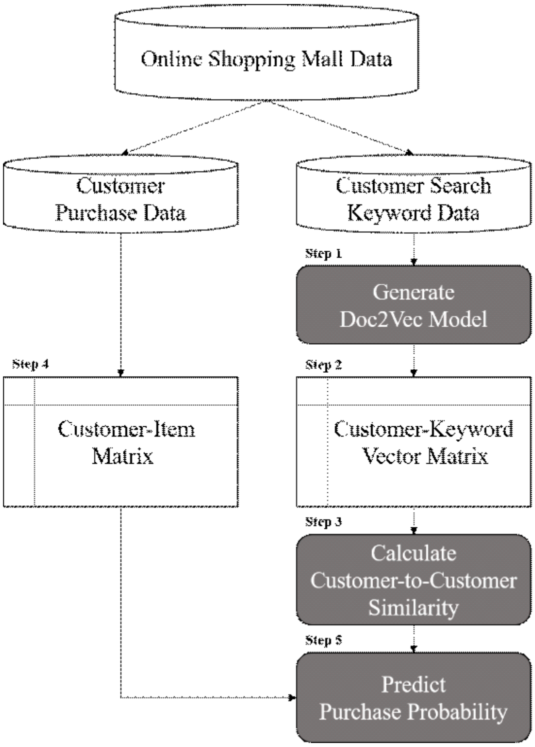
Research on recommender systems that focused on customer search behavior have been conducted several times in Korea, as well as abroad. Some studies have analyzed search behavior as only one type of online behavior (Jung *et al.*, 2012; Wu and Yan, 2017). On the other hand, other research has been conducted focusing only on search behavior (Cho and Lim, 2019; Kim *et al.*, 2019; Lee and Kim, 2019).

First, one study that analyzed customer search behavior as only one type of online behavior suggested a recommender system based on online bookstore data by utilizing online customer behavior and solved the conventional CF cold start problem (Jung *et al.*, 2012). Another study modeled the online behavior of customers by analyzing the customer's online behavior information by session (Wu and Yan, 2017). In both studies, the performance of the recommender systems was improved by the customer's online behavior information, including their search behavior. Recommender system research conducted mainly on customer search behavior also found that search behavior reflects the preferences of customers while showing improved recommendation performance results. Cho and Lim (2019) classified customer search keywords and proposed a hybrid CF. Each product was classified according to the keyword searched for and a product with high search keyword relevance was then recommended to new customers. Kim *et al.* (2019), who proposed a hybrid recommendation system, created a cluster of customers by processing the shopping information of groups that searched for similar keywords based on the customer's search history. They demonstrated that the results prevented the traditional CF cold start problem.

As such, studies on recommender systems based on customer search behavior can be found throughout domestic and foreign research. The results of recommender system research that includes search behavior shows that search behavior can be a factor that reflects customer preferences and has a positive effect on recommender system performance in that performance is improved over conventional recommendation systems. However, the number of cases and studies of recommender systems using customer search behavior in the e-commerce market is still quite small. In addition, even in the search behavior-oriented recommender system research, the studies are limited to keyword classification or recommender systems for customer clustering. So, it is necessary to pay attention to this fact because research on a more personalized recommendation system has not yet been sufficiently conducted.

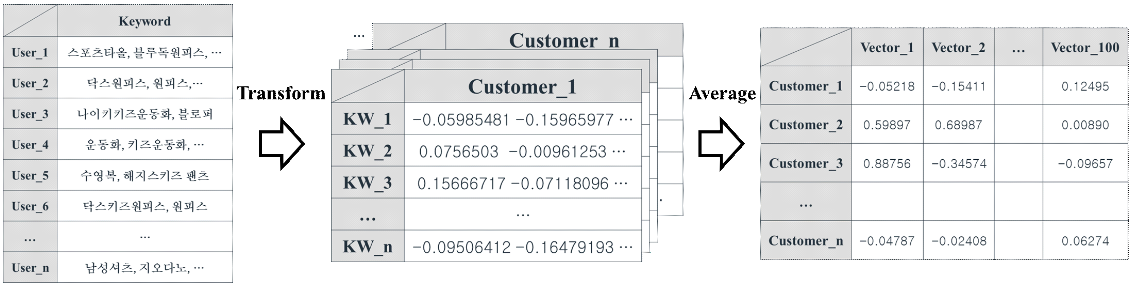
3. Suggested Model

This research was conducted to compare the quality of user-based CF that derives similarities between customers through customer purchase history data using the feature of Doc2Vec-based CF that derives similarities between customers through customer search keyword data. The proposed model is illustrated in <Figure 1>.


Online Shopping Mall Data
Customer Purchase Data
Customer Search Keyword Data
Step 1
Generate Doc2Vec Model
Step 4
Step 2
Customer-Item Matrix
Customer-Keyword Vector Matrix
Step 3
Calculate Customer-to-Customer Similarity
Step 5
Predict Purchase Probability

〈Figure 1〉 Suggested Model

(Step 1) First, all the search keywords of customers who have purchased at least *m* items among the top *N* items sold are extracted. The search keywords of each individual customer are considered as one document. Then, Doc2Vec is applied to the documents, which consist of customer's search keywords. This was influenced by the model proposed by Phi *et al.* (2016) and is



	Keyword
User_1	스포츠타올, 블루독원피스, …
User_2	닥스원피스, 원피스,…
User_3	나이키키즈운동화, 블로퍼
User_4	운동화, 키즈운동화, …
User_5	수영복, 해지스키즈 팬츠
User_6	닥스키즈원피스, 원피스
…	…
User_n	남성셔츠, 지오다노, …

	Customer_1
KW_1	−0.05985481 −0.15965977 …
KW_2	0.0756503 −0.00961253 …
KW_3	0.15666717 −0.07118096 …
…	…
KW_n	−0.09506412 −0.16479193 …

	Vector_1	Vector_2	…	Vector_100
Customer_1	−0.05218	−0.15411		0.12495
Customer_2	0.59897	0.68987		0.00890
Customer_3	0.88756	−0.34574		−0.09657
…				
Customer_n	−0.04787	−0.02408		0.06274

〈Figure 2〉 Representation of Customers and Search Keywords

explained in <Figure 2>.

<Figure 2> describes the process of transforming customer search keyword data to Doc2Vec. Search keyword data was pre-processed in the form of documents by combining all search keywords for each customer. Then the keyword document searched by each customer was converted into a vector using Doc2Vec. The resulting keyword vector is averaged to form a customer-keyword matrix.

Doc2Vec is an extended model of Word2Vec. Specifically, it assigns coordinate values for each word to a multi-dimensional space called the semantic space. It then trains the coordinates of the document along with the word. An individual customer's search keywords, which is considered as a document, is trained using Doc2Vec. Then the vector values of all the keywords are averaged to calculate the preference vector values for each customer.

(Step 2) The Customer-Keyword Vector Matrix is built using the vector values calculated in Step 1.

(Step 3) Similarity between customers is obtained by calculating Euclidean distances (*see* Equation (1)) using the vector values for each customer.

$$S(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

(Step 4 & Step 5) Then, the probabilities of purchasing each product are predicted based on CF using the Doc2Vec-based similarity between customers, which was derived in Step 3. At this time, our model makes a prediction by applying Equation (2), where $P_{x,I}$ is one (1) when customer *x* purchases product *i*, zero (0) otherwise. Thus, the customer-item matrix that contains the values of $P_{x,i}$ should be prepared before making purchase predictions.

$$P_{x,i} = \bar{P}_x + \frac{\sum_{y \in N} (P_{y,i} - P_y) \cdot S(x, y)}{\sum_{y \in N} |S(x, y)|} \tag{2}$$

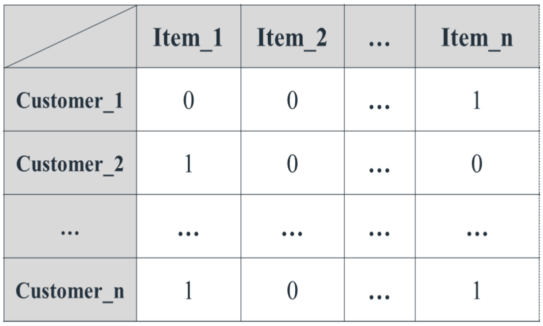
where is the average purchase prediction probability of customer *x* and *S(x,y)* is the similarity between the recommended customer *x* and neighbor customer *y*. *N* denotes a set of purchasers and *y* denotes an index indicating each neighbor.

4. Empirical Analysis

The data used in this research were provided from the 5th L.POINT Big Data Competition hosted by Lotte Members, L.POINT (Lotte Members, 2019). The L.POINT Big Data Competition is Korea’s representative big data competition that analyzes big data and develops content that is suitable for the theme based on actual data provided by Lotte’s integrated membership service. Lotte Members has approximately 36 million members, more than 60% of whom are South Koreans, and has a vast amount of lifestyle data as an integrated membership brand that combines Lotte’s 50 subsidiary companies as well as their external partners. In this study, customers’ purchase histories, customers’ search keywords, membership information, session information, and product classification data were used.

First, we picked the top 50 items most often sold and extracted 187 customers who bought five or more items among these top 50 items. Then, we extracted all the search keywords for every one of the 187 customers. The search keywords for each individual customer were considered as one document and each customer was mapped into the vector space using Doc2Vec. Euclidean distances were used to calculate the similarity between customers. Finally, the model predicted the purchase probabilities for every item for each customer using Equation (2).

To compare it with the proposed model, we generated a conventional CF model that only used customers’ purchase history data. For this comparison model, the customer-item matrix built using the purchase history ($P_{x,i}$) of 187 customers was used to calculate the similarity between customers. <Figure 3> shows an example of the customer-item matrix used in this study.



	Item_1	Item_2	…	Item_n
Customer_1	0	0	…	1
Customer_2	1	0	…	0
…	…	…	…	…
Customer_n	1	0	…	1

〈Figure 3〉 An Example of a Customer-Item Matrix

As shown in <Figure 3>, the elements of the customer-item matrix are binary-valued (1 or 0), which indicates whether a customer purchased a product (1) or not (0). The similarity between the customers was calculated using Jaccard similarity (3) from the derived customer-item matrix. The Jaccard similarity has a value ranging from zero (0) to one (1). It equals one (1) when the purchases are identical between the two customers and zero (0) when the purchases between the two customers completely differ.

$$\mathrm{sim}(x,y) = \frac{M_{11}}{M_{10} + M_{01} + M_{11}} \tag{3}$$

M_{11} : 1 if customer x and y both purchased, 0 otherwise
M_{10} : 1 if customer x only purchased, 0 otherwise
M_{01} : 1 if customer y only purchased, 0 otherwise

To compare the performance of the proposed CF and the conventional CF model, we used F1-measure (4) as a performance measure. It is a harmonic mean of recall and precision with the same weight as an index for measuring performance. Here, recall (5) is defined as the ratio of the products recommended by the CF algorithm among the products actually purchased by the target customer and precision (6) is the ratio of the product that a customer actually purchased among the products recommended by the CF algorithm (Sarwar *et al.*, 2000).

$$F1 - measure = \frac{2 \times recall \times precision}{recall + precision} \tag{4}$$

$$Recall = \frac{Products\ purchased\ by\ the\ customer \cap Recommended\ products}{Products\ purchased\ by\ the\ customer} \tag{5}$$

$$\mathrm{Precision} = \frac{Products\ purchased\ by\ the\ customer \cap Recommended\ products}{Recommended\ products} \tag{6}$$

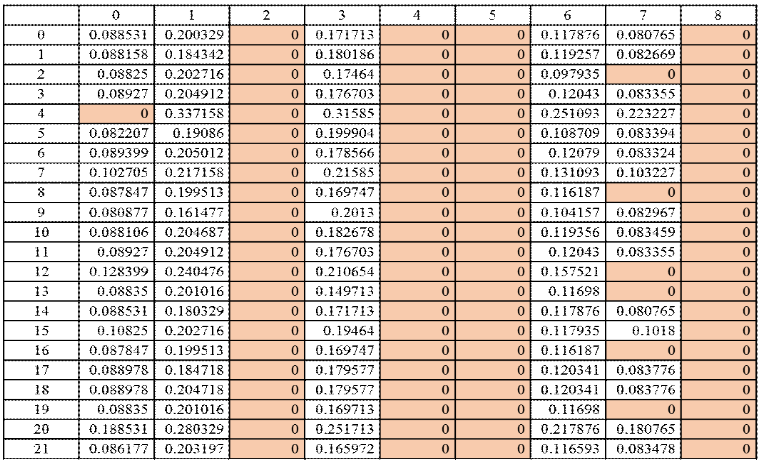
The experimental results are presented in <Table 1>. We compared F1-scores for the top 3, 5, and 7 items recommended by the two models. The performance of the conventional CF was highest when the number of recommended products was the smallest (*i.e.*, the Top 3). But as the number of recommended products increased, the F1-scores continued to decrease. On the other hand, the performance of the proposed model (*i.e.*, the hybrid CF model based on the customer's search keywords) increased as the number of recommended products increased. In the two cases (*i.e.*, Top 5 and Top 7), the proposed CF outperformed the conventional CF. This indicates that customer search keywords can sufficiently reflect customer preferences.

In general, a customer collects information about a product through a number of search behaviors before purchasing the product and, in the process, generates a large amount of search keyword data. As the number of recommended products increases, the performance of the proposed CF model increases because search keyword data is less scarce than purchase data and

〈Table 1〉 F1-Scores of the models

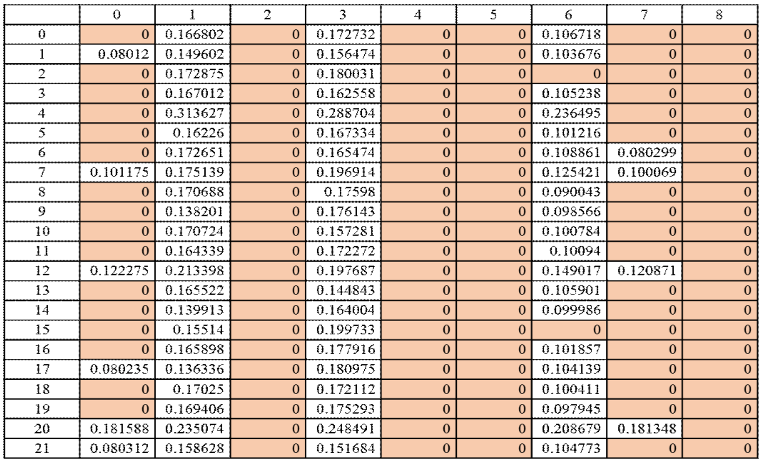
Model	The Number of Recommended Products	Precision	Recall	F1-Score
Purchase-Based CF	Top3	0.258	0.165	0.199
	Top5	0.176	0.187	0.179

customer preferences are well reflected therein. The conventional CF performance based only on purchasing data that indicated a strong customer preference was highest when there were three recommended products, but its performance decreased as the number of recommended products increased due to the scarcity problem. However, the proposed CF model is able to recommend many products because it is able to address the scarcity problem—even when the number of recommended products increases. This is explained in <Figure 4>. <Figure 4> is a part of the customer-item matrix of the two models. It demonstrates the fact that the conventional CF matrix is sparser than the proposed CF model matrix.



	0	1	2	3	4	5	6	7	8
0	0.088531	0.200329	0	0.171713	0	0	0.117876	0.080765	0
1	0.088158	0.184342	0	0.180186	0	0	0.119257	0.082669	0
2	0.08825	0.202716	0	0.17464	0	0	0.097935	0	0
3	0.08927	0.204912	0	0.176703	0	0	0.12043	0.083355	0
4	0	0.337158	0	0.31585	0	0	0.251093	0.223227	0
5	0.082207	0.19086	0	0.199904	0	0	0.108709	0.083394	0
6	0.089399	0.205012	0	0.178566	0	0	0.12079	0.083324	0
7	0.102705	0.217158	0	0.21585	0	0	0.131093	0.103227	0
8	0.087847	0.199513	0	0.169747	0	0	0.116187	0	0
9	0.080877	0.161477	0	0.2013	0	0	0.104157	0.082967	0
10	0.088106	0.204687	0	0.182678	0	0	0.119356	0.083459	0
11	0.08927	0.204912	0	0.176703	0	0	0.12043	0.083355	0
12	0.128399	0.240476	0	0.210654	0	0	0.157521	0	0
13	0.08835	0.201016	0	0.149713	0	0	0.11698	0	0
14	0.088531	0.180329	0	0.171713	0	0	0.117876	0.080765	0
15	0.10825	0.202716	0	0.19464	0	0	0.117935	0.1018	0
16	0.087847	0.199513	0	0.169747	0	0	0.116187	0	0
17	0.088978	0.184718	0	0.179577	0	0	0.120341	0.083776	0
18	0.088978	0.204718	0	0.179577	0	0	0.120341	0.083776	0
19	0.08835	0.201016	0	0.169713	0	0	0.11698	0	0
20	0.188531	0.280329	0	0.251713	0	0	0.217876	0.180765	0
21	0.086177	0.203197	0	0.165972	0	0	0.116593	0.083478	0

(a) Customer-item matrix of the proposed model



	0	1	2	3	4	5	6	7	8
0	0	0.166802	0	0.172732	0	0	0.106718	0	0
1	0.08012	0.149602	0	0.156474	0	0	0.103676	0	0
2	0	0.172875	0	0.180031	0	0	0	0	0
3	0	0.167012	0	0.162558	0	0	0.105238	0	0
4	0	0.313627	0	0.288704	0	0	0.236495	0	0
5	0	0.16226	0	0.167334	0	0	0.101216	0	0
6	0	0.172651	0	0.165474	0	0	0.108861	0.080299	0
7	0.101175	0.175139	0	0.196914	0	0	0.125421	0.100069	0
8	0	0.170688	0	0.17598	0	0	0.090043	0	0
9	0	0.138201	0	0.176143	0	0	0.098566	0	0
10	0	0.170724	0	0.157281	0	0	0.100784	0	0
11	0	0.164339	0	0.172272	0	0	0.10094	0	0
12	0.122275	0.213398	0	0.197687	0	0	0.149017	0.120871	0
13	0	0.165522	0	0.144843	0	0	0.105901	0	0
14	0	0.139913	0	0.164004	0	0	0.099986	0	0
15	0	0.15514	0	0.199733	0	0	0	0	0
16	0	0.165898	0	0.177916	0	0	0.101857	0	0
17	0.080235	0.136336	0	0.180975	0	0	0.104139	0	0
18	0	0.17025	0	0.172112	0	0	0.100411	0	0
19	0	0.169406	0	0.175293	0	0	0.097945	0	0
20	0.181588	0.235074	0	0.248491	0	0	0.208679	0.181348	0
21	0.080312	0.158628	0	0.151684	0	0	0.104773	0	0

(b) Customer-item matrix of the conventional model

5. Conclusion

This research proposed a novel hybrid CF that is designed to utilize customers' search keywords data. The model proposed in this study solved the scarcity problem encountered by recommender systems based on search keyword data and not only found customer preferences through the search data but also proposed a new recommendation system using a word embedding technique.

The existing recommendation system research has been conducted using ratings data. However, since the recommendation system is applied in various ways without any boundaries between industries and sectors, the data generated naturally by the company's value activities should also be used as recommendation system data. However, if the recommendation system is generated using only purchase history data, encountering the sparsity problem due to a lack of information is inevitable. Customer search keyword data continues to increase as the customer performs each search action and, therefore, more information is generated over time. In addition, unlike customer reviews, search keywords are composed of nouns, so data pre-processing is less difficult and the amount of information lost during pre-processing is small. Therefore, this study used search keyword data to address the lack of information in purchase history and ratings data.

Based on previous research that indicated customer preferences through search keywords (Gang and Choi, 2019), an empirical analysis of this study's proposed model demonstrated that as the number of recommended products increased, the proposed model performed better than the conventional CF model. Through these results, it was proven that search keywords can represent customer preferences and that, if customer search keyword data are similar to each other, recommendations are possible. This has practical significance in that the industry can use search keyword data to grasp customer preferences and recommend products. In addition, it has academic significance in that it proposes a new hybrid recommendation system by combining Doc2Vec—which was previously used only as a document embedding and document classification tool for domestic research—with a recommendation system.

However, our study has some limitations. First, common purchase patterns across customers were insufficient since the data used in the empirical analysis were sampled. Therefore, we need to extend the amount of experimental data in future studies. In addition, verifying the proposed model using a dataset other than Lotte's is required.

Secondly, when modeling with Doc2Vec, we did not consider repeated occurrences of search keywords. It is important to pay attention to repeated search keywords because when search keywords are repeated several times it reflects a strong customer preference.

In addition, in order to enable better performance, an effort to develop a more proper document embedding method should be undertaken in the future. Therefore, it is necessary to attempt additional various word embedding techniques in future studies.

References

Adomavicius, G., and A. Tuzhilin, "Context-aware recommender systems," *In Recommender Systems Handbook,* Springer, Boston, MA, 2011

Balabanović, M., and Y. Shoham, "Fab: content-based, collaborative recommendation," *Communications of the ACM*, Vol.40, No.3 (1997), 66~72.

Billsus, D., and M. J. Pazzani, "Learning Collaborative Information Filters," *In International Conference of Machine Learning,* Vol. 98, (1998), 46~54.

Breese, J. S., D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, (1998), 43~52.

Cho S.-e., and H.-s. Lim, "A Study on Product Recommendation System Based on User Search keyword," *Journal of Digital Contents Society*, Vol.20, No.2(2019), 315~320.

Cho, S.-y., J.-e. Choi, K.-h. Lee, and H.-w. Kim, "An Online Review Mining Approach to a Recommendation System," *Information Systems Review*, Vol.17, No.3(2015), 95~111.

Cho, Y. H., J. K. Kim, D. H. Ahn, and H. A. Lee, "An Explanation based Recommender System using Collaborative Filtering: WebCF-Exp," *Korean Management Review*, Vol.35, No.2 (2006), 493~19.

Choi D.-j., J.-y. Park, S.-b. Park, J.-t. Lim, J.-o Song, K.-s. Bok, and J.-s. Yoo, "Personalized Recommendation Considering Item Confidence in E-Commerce," *Journal of the Korea Contents Association*, Vol.19, No.3(2019), 171~182.

Choi, S. B. and H. C. Ahn, "A study on the improvement of collaborative filtering prediction accuracy using online user review sentiment analysis," *In Proceedings of the Korea Intelligent Information System Society Conference*, (2017), 30~31.

Kang, E. J. and Y. S. Choi, "Influence of information retrieval using A.I speaker on online purchasing experience - based on AISAS model," In *Proceedings of HCI Korea 2019*, (2019), 425~430

Karvelis, P., D., Gavrilis, G., Georgoulas, and C. Stylios, "Topic recommendation using Doc2Vec," *In 2018 IEEE International Joint Conference on Neural Networks (IJCNN*), (2018), 1~6.

Ki, H., J.-h. Lee, H.-w. Park, M.-j. Chae, S.-w. Choi and J. Park, "Inferring User Traits from Applications Installed on a Smart Phone," *Journal of KIISE: Software and Applications*, Vol.45, No.12(2018), 1240~1249.

Jeong, J., M. Jee, M. Go, H. Kim, H. Lim, Y. Lee, and W. Kim, "Related Documents Classification System by Similarity between Documents," *Journal of Broadcast Engineering*, Vol.24, No.1(2019), 77~86.

Jeong J.-w., W.-s. Hwang, H.-j. Lee, S.-w. Kim, "Recommendation Exploiting Search-Keywords in Online Shopping," *Proceedings on KIISE Conference*. Vol. 39. No. 2(2012), 95~97

Kim, H., S. Chae, J. Yoo, and S. Bae, "Online shopping mall customer purchasing type clustering and purchasing power evaluation model," *Proceedings on Korean Institute of Industrial Engineers*, 2019, 2597~2616

Kim, K. S., "A hybrid collaborative filtering algorithm for personalized recommendations and its application to the internet electronic commerce," *The Journal of Internet Electronic Commerce Research,* Vol.8, No.4(2008), 1~20.

Kim, M. G. and K.-j. Kim, "Recommender Systems using Structural Hole and Collaborative Filtering," *Journal of Intelligence and Information Systems*, Vol.20, No.4(2014), 107~120.

Kim, M., N. G. Kim, and I. H. Jung, "A Methodology for Extracting Shopping-Related Keywords by Analyzing Internet Navigation Patterns," *Journal of Intelligence and Information Systems*, Vol.20, No.2(2014), 123~136.

Ku, M. J. and H. C. Ahn, "A Hybrid Recommender System based on Collaborative Filtering with Selective Use of Overall and Multicriteria Ratings," *Journal of Intelligence and Information Systems*, Vol.24, No.2(2018), 85~109.

Le, Q. and T., Mikolov, "Distributed representations of sentences and documents," *In Proceedings of the International Conference on Machine Learning,* (2014), 1188-1196.

Lee, H. S. and P. S. Kim, "The Effect of Consumer's Technology Acceptance and Resistance on Intention to Use of Artificial Intelligence (AI)," *Korean Management Review*, Vol.48, No.5(2019), 1195~1219

Lee, R. K., N. H. Chung, and T. H. Hong, "Developing the online reviews based recommender models for multi-attributes using deep learning," *Journal of Information Systems*, Vol.28, No.1(2019), 97~114.

Lee Y. S., K. C. Cha, and S.-h. Kim. "Internet Search Behavior and Box Office Performance," *Korean Management Review*, Vol.45, No.5 (2016), 1501~1526.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, (2013), 3111~3119.

Nandi, R. N., M. A., Zaman, T., Al Muntasir, S. H., Sumit, T., Sourov, and M. J. U. Rahman, "Bangla news recommendation using doc2vec," *In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP),* (2018), 1~5.

Park, J., Y. H. Cho, and J. K. Kim, "Social Network : A Novel Approach to New Customer Recommendations," *Journal of Intelligence and Information Systems*, Vol.15, No.1(2009), 123~140.

Park, J., and N. Kim, "Multi-Vector Document Embedding Using Semantic Decomposition of Complex Documents," *Journal of Intelligence and Information Systems*, Vol.25, No.3 (2019), 19~41.

Phi, V. T., L., Chen, and Y. Hirate, "Distributed representation based recommender systems in e-commerce," *In DEIM Forum*, (2016).

Sarwar, B., G., Karypis, J. A. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for E-commerce," *Proceedings of ACM E-commerce 2000 Conference*, (2000), 158~167.

Shim, J., H. R. Won, and H. Ahn, "Doc2vec-based intelligent fake news classification model using domestic news articles," *Proceedings of* *the Korea Intelligent Information System Society Conference*, 1~3.

Shin J. H., J. H. Song, K. S. Bok and J. S. Yoo, "Personalized Travel Destination Recommendation Scheme through Hybrid Collaborative Filtering," *Proceedings of the Korea Contents Association*, (2018), 383~384.

Stiebellehner, S., J., Wang, and S., Yuan, "Learning Continuous User Representations through Hybrid Filtering with doc2vec," *arXiv preprint arXiv 1801.00215*., (2017).

Takács, G., I., Pilászy, B. Németh, and D., Tikk, "Scalable collaborative filtering approaches for large recommender systems." *Journal of Machine Learning Research,* (2009), 623~656.

Wu, C., and M. Yan, "Session-aware information embedding for e-commerce product recommendation," *In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, (2017), 2379~2382.

Zhang, X., J., Liu, B., Cao, Q., Xiao, and Y. Wen, "Web Service Recommendation via Combining Doc2Vec-Based Functionality Clustering and DeepFM-Based Score Prediction," *In 2018 IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/Social Com/SustainCom),* (2018), 509~516.

국문요약

검색 키워드를 활용한 하이브리드 협업필터링 기반 상품 추천 시스템

이윤주* · 원하람* · 심재승* · 안현철**

추천시스템(recommender system)은 고객의 선호도를 예측하여 상품과 서비스를 제공하는 기법으로, 현재 다양한 온라인 서비스에 활용되고 있다. 이와 관련된 많은 선행 연구들은 협업필터링(collaborative filtering)에 기반한 추천시스템을 제안하였는데, 대부분의 경우 고객의 구매 내역 또는 평점 데이터만 사용하여 진행되었다. 오늘날 소비자들은 제품을 구매하는 과정에서 온라인 검색 행동을 하여 관심있는 제품을 찾는다. 그렇기 때문에 검색 키워드 데이터는 고객의 선호도를 파악하는데 매우 유용한 정보일 수 있다. 그러나 지금까지 추천시스템 연구에서 사용되는 경우는 거의 없었다. 이에 본 연구는 고객의 검색 행동에 주목하여 온라인 쇼핑몰 고객의 검색 키워드 데이터와 구매 데이터를 고려한 하이브리드 협업 필터링을 제안하였다. 본 연구는 제안된 모델의 적용 가능성을 검증하기 위해 실제 온라인 쇼핑몰 데이터를 사용하여 성능을 검증하였다. 연구 결과, 추천 상품의 개수가 많아질수록 고객의 검색 키워드를 기반으로 구축된 협업필터링의 추천 성능이 향상되는 반면 일반적인 협업필터링의 성능은 추천된 상품의 개수가 많아질수록 점차 감소함을 발견하였다. 따라서 본 연구는 검색 키워드 데이터를 활용한 하이브리드 협업필터링이 고객의 선호도를 반영한 추천할 수 있으며, 구매이력 데이터의 정보부족을 해결할 수 있음을 확인하였다. 이는 기존의 정량 데이터만을 활용한 추천 시스템이 아닌, 비정형 데이터인 텍스트를 사용함으로써 새로운 하이브리드 협업필터링 구축 방법을 제안했다는 점에서 의의가 있다.

주제어 : 추천시스템, 협업필터링, Doc2Vec, 워드 임베딩, 검색 키워드

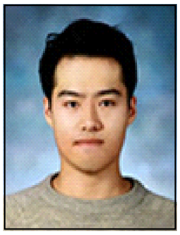
논문접수일 : 2020년 2월 3일 논문수정일 : 2020년 3월 21일 게재확정일 : 2020년 3월 24일
원고유형 : 단편논문 교신저자 : 안현철

* 국민대학교 비즈니스 IT 전문대학원
** Corresponding Author: Hyunchul Ahn
Graduate School of Business IT, Kookmin University,
77 Jeongneung-ro, Seongbuk-gu, Seoul, 02707, Republic of Korea
Tel: +82-2-910-4577, Fax: +82-2-910-4017, E-mail: hcahn@kookmin.ac.kr

저자소개

Yunju Lee

She holds a bachelor's degree from Kangwon National University, Korea and is currently a master's degree in Business Analysis Track at Business IT Graduate School, Kookmin University in Korea. Her interests include business analytics, data mining and marketing.



Haram Won

He holds a bachelor's degree in business administration from Halla University, Korea, and is currently a master's degree in Business Analysis Track at Business IT Graduate School, Kookmin University in Korea. His interests include business analytics and CRM and data-driven marketing analytics.

Jaeseung Shim

He is currently a master's program at Graduate School of Business IT, Kookmin University, where he earned his bachelor degree in Management Information. His primary research interests include data mining and machine learning for the social sciences and business.

Hyunchul Ahn

He is a professor of management information systems at the Graduate School of Business IT, Kookmin University, Republic of Korea. He obtained his Ph.D. from Korea Advanced Institute of Science and Technology. His major research interests are intelligent IS and IS adoption. His research has been published in *Annals of OR, Applied Intelligence, Computers & OR, Expert Systems with Applications, International Journal of Electronic Commerce, International Journal of Production Research, Information & Management*, etc.