# A Hybrid Collaborative Filtering Model Using Customer Search Keyword Data for Product Recommendation

Ha-Ram Won
Graduate School of Business IT
Kookmin University
Seoul, Republic of Korea
haramy44@kookmin.ac.kr

Yunju Lee
Graduate School of Business IT
Kookmin University
Seoul, Republic of Korea
mmlas0ui@kookmin.ac.kr

Jae-Seung Shim
Graduate School of Business IT
Kookmin University
Seoul, Republic of Korea
simsoni7@kookmin.ac.kr

Hyunchul Ahn
Graduate School of Business IT
Kookmin University
Seoul, Republic of Korea
hcahn@kookmin.ac.kr

***Abstract*— 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) has been used the most as an algorithm for implementing recommender systems. However, in most cases, it has only used purchase history or customer's ratings though there are numerous available data provided by customers. E-commerce customers frequently use a search function to find the products they are interested in among the vast array of products offered. Such search keywords data may be a very useful information source for modeling customer's preference. 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 Doc2Vec 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 a real-world online shopping mall data in Korea. As a result, we found that search keywords data might effectively represent the preference of customers, and contribute to the improvement of conventional CF.**

**Keywords— Recommender System, Hybrid Collaborative Filtering, Doc2Vec, Search Keyword**

## I. INTRODUCTION

Recommender systems, which are widely used today in 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 prefers or needs [1] [2]. Especially for online businesses that have a lot of customers and product data, a sophisticated recommender system that accurately predicts customer preferences can be a competitive advantage since it may lead to better customer satisfaction and sales increase [3].

Collaborative Filtering (CF), which has been evaluated as a high-performance recommendation algorithm in the industry and academia, is a method of recommending to customers the products preferred by using the similarity between users or items [2] [3] [4]. There are two types of CF: user-based or item-based recommendations. Item-based CF predicts the rating of a user for an item based on the ratings of the user for the items similar to the target item [5]. In this type of CF, the similarity between users is not considered. Thus, its recommendation quality may be lower than user-based CF when users have different preferences.

On the other hand, user-based CF predicts the preference of a user for an item using the ratings for this item by other users that have similar rating patterns. In detail, it analyzes the similarity between users to find the users most similar to the target user, and recommends the product to the target user based on the purchase history or ratings of his/her similar users (which is called 'neighbors').

Though CF is one of the most popular recommendation algorithms, it is criticized because of its several problems. First, when the users have seldom purchased or the items have been seldom sold (i.e. when the users or the items are new), cold-start problem occurs. In this case, it becomes difficult to infer ratings of similar users or items. Second, when the total number of products is much larger than the number of products that a customer purchases, sparsity problem may occurs. In this case, it becomes highly probable that the similarity between users or items becomes zero, which makes CF useless. Third, most CF studies have only used customer's purchase history or ratings when measuring similarity between users or items even though there are numerous available data provided by customers [6] [7] [8].

To mitigate these problems, many researchers have proposed hybrid CF. For example, Shin et al. [9] proposed a personalized travel destination recommendation based on hybrid filtering and predicted the expected scores of the destinations that travelers did not visit, and [10] performed additional rating prediction by combining similar user ratings and similar product ratings with a hybrid CF to solve the lack of data.

Most prior studies have suggested recommendation system using only structured data such as ratings or purchase history. But, in the field of online services that recommendation system is actively utilized, there are also various unstructured data such as customers' reviews and search keywords. Reference [11] proposed hybrid filtering model for app users based on document embedding. They collected app history and app description data from the Apple's App store and Google's Play store, and analyzed these unstructured data using Doc2Vec technique. They derived conclusion that the quality of recommendations was improved when combined with CF. With such backgrounds,

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00251

this research tries to compare the quality of the hybrid CF combined with Doc2Vec using structured data(purchase history data) and unstructured data(customer search keyword data) of online shopping malls with conventional CF using only structured data.

## II. Relarted Research

### A. Doc2vec

Doc2Vec is an extended method of Word2Vec that finds associations between words in sentences and converts them into vectors. Doc2Vec make a comparison between documents, using a simple artificial neural network to find similar documents, and then locate them closely on multidimensional spaces [12]. 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) [13]. First, Distributed Memory method, which is similar to CBOW (Continuous Bag of Words) method of Word2Vec combines the document ID vector and the word vectors of the document to predict other words in the document. In DM method, a word vector inherits the meaning of the word and has the same meaning in all documents. Distributed Bag of Words 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 Skip-Gram method of Word2Vec [13].

### B. Search Keyword

This research focused on customer search keyword data on online shopping mall as unstructured data to be used for the model. According to [14], the customer's search behavior is an act of collecting information on the products of interest, which may indicate the customer's preference. Therefore, the customer's search behavior can influence the buying behavior of the customer, and it can be used for recommending products to a similar customer based on the purchase history of the customer whose search keywords are similar.

## III. Suggested Model

This study aims to propose a novel hybrid user-based CF model which derives the similarities between customers using customer's search keyword data. The proposed model is shown in Fig. 1.

(Step 1) First, all search keywords of customers who have purchased at least $m$ items among the top $N$ items sold are extracted, and then search keywords of an individual customer are considered as one document. And then, Doc2Vec is applied to the documents which are consist of customer's search keywords.

Doc2Vec assigns coordinate values of each word to a multidimensional space called semantic space, and then it trains the coordinates of the document along with the word. The search keywords of an individual customer, which is considered as a document, is trained using Doc2Vec, and then the vector values of all the keywords is averaged to calculate the preference vector values for each customer.

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

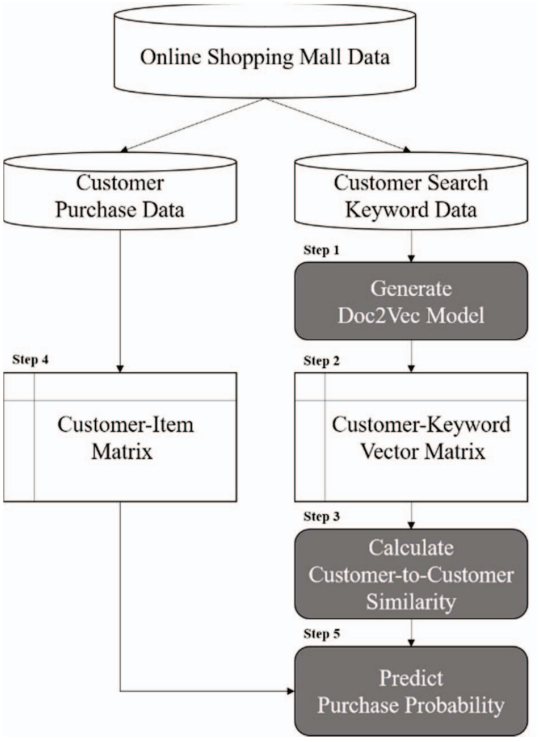


Fig. 1. Suggested Model

(Step 3) The similarity between the customers is obtained by using the Euclidean distance (1) using the vector value of 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 the customers, which is derived in Step 3. At this time, our model makes a prediction by applying equation (2) where $P_{x,i}$ is 1 when customer $x$ purchases product $i$, 0 otherwise. Thus, the customer-item matrix which 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 $\overline{P_x}$ is the average purchase prediction probability of customer $x$, and $S(x,y)$ is the similarity between the recommended customer $x$ and the neighbor customer $y$. $N$ denotes a set of purchasers and $y$ denotes an index indicating each neighbor.

## IV. Empirical Analysis

The data used in this research was provided from the 5th L.POINT Big Data Competition hosted by Lotte Members, L.POINT [15]. 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

| | Item_1 | Item_2 | … | Item_n |
|---|---|---|---|---|
| User_1 | 0 | 0 | … | 1 |
| User_2 | 1 | 0 | … | 0 |
| … | … | … | … | … |
| User_n | 1 | 0 | … | 1 |

Fig. 2. An Example of Customer-Item Matrix

data provided by Lotte's integrated membership service. Lotte Members has approximately 36 million members, more than 60% of South Koreans, and has a vast amount of lifestyle data as an integrated membership brand that combines Lotte's fifty subsidiary companies and their external partners. In this research, customers' purchase history, customers' search keywords, membership information, session information, and product classification data were used.

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

To compare it with the proposed model, we generated a conventional CF model which only uses customer's purchase history data. On this comparison model, the customer-item matrix built using the purchase history ($P_{x,i}$) of 187 customers is used for calculating the similarity between customers. Fig. 2 shows an example of customer-item matrix used in this study.

As shown in Fig. 2, the elements of the customer-product 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. Jaccard similarity has a value ranging from 0 to 1. It becomes 1 when the purchases are identical between the customers and becomes 0 when the purchases between the two customers are completely different.

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

$M_{11}$ : 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 performances between the proposed CF and conventional CF models, 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 recommended products by the CF algorithm among the products actually purchased by the target customer, and precision (6) as the ratio of the product a customer actually purchased among the recommended products by the CF algorithm [16].

TABLE I. F1-SCORES OF THE MODELS

| | Top3 | Top5 | Top7 |
|---|---|---|---|
| **Purchase-based CF (Conventional CF)** | 0.199 | 0.179 | 0.175 |
| **Keywords-based CF (Proposed CF)** | 0.179 | 0.195 | 0.193 |

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

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

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

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

## V. CONCLUSION

This research proposed a novel hybrid CF that is designed to utilize customer's search keywords data. Prior studies on CF have seldom used it because they are unstructured and hard to be handled. Most of them have just used customer's ratings or purchase history. Using ratings or purchase history often leads to sparsity problem.

In order to mitigate this problem, this research proposed a method for constructing hybrid CF model through text analysis of search keywords data that can represent customer's preference. Using a real-world dataset, we validated its usefulness by an empirical analysis.

However, our study has some limitations. First, the common purchase patterns across the customers were insufficient since the data used in the empirical analysis was the sampled one. Second, when modeling Doc2Vec, we did not consider the repeated occurrences of the search keywords. Therefore, we need to extend the size of the experimental data in future study. In addition, the efforts for developing a more proper document embedding method should be done in the future. Finally, it is also needed to verify the proposed model using the dataset other than Lotte's case.

## REFERENCES


[1] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Scalable collaborative filtering approaches for large recommender systems," *Journal of machine learning research*, 10(March):623-656, 2009.

[2] J. H. Park, Y. H. Cho, and J. K. Kim, "Social network : a novel approach to new customer recommendations," *Journal of Intelligence and Information Systems*, 15(1):123-140, 2009.

[3] M. Kim and K. Kim, "Recommender systems using structural hole and collaborative filtering," *Journal of Intelligence and Information Systems*, 20(4):107-120, 2014.

[4] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43-52, 1998.

[5] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Recommendation systems handbook, Springer, 2011.

[6] M. J. Ku, and H. Ahn, "A hybrid recommender system based on collaborative filtering with selective use of overall and multicriteria ratings," *Journal of Intelligence and Information Systems*, 24(2):85-109, 2018.

[7] K. Kim, "A hybrid collaborative filtering algorithm for personalized recommendations and its application to the internet electronic commerce," *The Journal of Internet Electronic Commerce Resarch,* 8(4):1-20, 2008,.

[8] J. K. Kim, D. H. Ahn, and Y. H. Cho, "A personalized recommender system, WebCF-PT: a collaborative filtering using web mining and product taxonomy," *The Journal of MIS Research*, 15(1):63-79, 2005.

[9] J. Shin, J. Song, K. Bok, and J. Yoo, "Personalized travel destination recommendation scheme through hybrid collaborative filtering," In *Proceedings of the Conference of the Korea Contents Society*, pages 383-384, 2018.

[10] D. Choi, J. Park, S. Park, J. Lim, J.-O. Song, K. Bok, and J. Yoo, "Personalized recommendation considering item confidence in e-commerce," *The Journal of the Korea Contents Association*, 19(3): 171-182, 2019.

[11] S. Stiebellehner, J. Wang, and S. Yuan, "Learning continuous user representations through hybrid filtering with Doc2vec," *arXiv preprint arXiv 1801.00215*, 2017.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," In *Advances in neural information processing systems*, pages 3111-3119, 2013.

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

[14] E. Kang and Y. Choi, "Influence of information retrieval using A.I speaker on online purchasing experience - based on AISAS model," In *Proceedings of HCI Korea 2019*, pages 425-430, 2019.

[15] Lotte Members, The 5th L.POINT big data competition. URL https://competition.lpoint.com/ (Accessed 30 August 2019).

[16] B. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Analysis of recommendation algorithms for e-commerce," In *Proceedings of ACM E-commerce 2000 conference*, pages 158-167, 2000.