APPLYING COLLABORATIVE FILTERING METHOD FOR DOCUMENT RECOMMENDER SYSTEM

Quoc-Hung Ly¹, T.Phuong-Nam Phan²*

Abstract – The recommender system helps recommend relevant information items to the user. In recommender systems, collaborative filtering is commonly used to gauge users' interest in new products. Collaborative filtering systems often rely on data about the similarity of users or products in the system in the past to predict preferences or new products for specific users. This study applies the collaborative filtering technique with the k-nearest neighbor to recommend documents for the English center. The implementation process encompasses the following sequential steps. Initially, a system is constructed to collect and store data in the database. Subsequently, the study implements a recommendation algorithm with three cases: Case 1: For new users; Case 2: For users who have seen the most document items; and Case 3: For center members. The results make it easier for users to find documents.

Keywords: collaborative filtering, k-nearest neighbor, predict, recommender system.

I. INTRODUCTION

Recommender systems (RS) have been applied in many different fields. For instance, firstly, in e-commerce, the RS is applied in sales websites such as Amazon, eBay, Lazada, Shopee, and Tiki. Secondly, in streaming video and music, RSs recommend video content, movies, TV shows, music, and playlists based on user preferences and viewing history, such as Netflix, Spotify, YouTube, and TikTok. Thirdly, in video games, the RS recommends similar games or games related to the user’s interests, for example, Steam, PlayStation, and Xbox. Next, the RS suggests friends, groups, and similar content with the user’s interests and activity history for social networks like Facebook, Instagram, LinkedIn, TikTok, and Twitter. Furthermore, in information search, the RSs are used to classify and search information for websites and applications, such as Google, Bing, Yahoo!, News, and magazines [1].

In addition, in the study of Tran Nguyen Minh Thu et al. [2], the RS supports searching documents to perform the proposed research using the collaborative filtering method of the RS in combination with the Elastic Search keyword indexing tool to improve the document search function. RS [3] applied to automatic news aggregator websites to help readers cope with the information explosion. Using content filtering and syndication methods, they built an RS that applies to an automated news aggregator website (NewsRES). Product RS in online sales using collaborative filtering techniques. This study used a collaborative filtering technique to evaluate user interest in new products [1]. This technique predicted user preferences for new products based on data on past product preferences of users. The neighborhood model was introduced for product recommendation in the online sales system and compared with other essential techniques to assess reliability Trieu Vinh Viem et al. [4] build a movie RS based on the neighbor factor model to present a web-based system to recommend movies for users and improve the prediction results by adjusting the model by regularization coefficients on each of the different parameters of the model.

Currently, there are many English centers in Tra Vinh Province, such as Victory Informatics - Foreign Language Center, UK English, iLingo Language School, ALOHA Saigon - Tra Vinh
English System, Viet - My Tra Vinh English Center, which has a website, which only introduces information about centers and courses, but not yet applied a RS to the website. There has yet to be a RS for a personal English center website. Building a RS on the website for an English center is entirely new, which can help to improve the user’s document search experience and web surfers.

In this study, the authors utilized a collaborative filtering technique combined with the k-nearest neighbor algorithm to implement document RS. The Pearson correlation similarity measure calculated the similarity between user u and u’.

II. RELATED WORKS AND RECOMMENDED CASES

A. Recommender system

Recommender systems are designed to provide users with relevant recommendations, the most effective recommendations possible from the information of data items, user profiles, and relationships between items [2]. A recommender system (RS) is a tool that provides valuable and personalized information on a system containing a large amount of information. A RS, a decision support system, suggests information related to users easily and quickly, suitable for each user [5].

A RS consists of three main components [5], such as the set of users \( U = u_1, \ldots, u_n \) including the user’s information stored on the system; the set of data items \( I = i_1, \ldots, i_m \) including the identifier and attributes of the data item; a set of relationships \( R = (U_{ui}, I_i) \) between users and data items. This is the set of associated transactions between a set of the U users with a set of the data I items and descriptions of this association. Figure 1 shows user-item-rating matrix.

In the RS, the feedback value of each user for the items will be recorded as a basis for predicting the following values. Depending on the system, this value will have different meanings. For example, it can be used to measure the relevance or likeness in e-commerce systems or the competence/performance of users in e-learning systems. The rui value can be specified through explicit feedback, such as rating. For example, rating from 1 star to 5 star; or like (1), and dislike (0). That means u voted for i, or rui can be determined by implicit feedback by the number of mouse clicks. That means u viewed i.

B. Collaborative filtering (CF)

Collaborative filtering emerged in the early 1990s as a response to the problem of information overload in online spaces. CF is a recommendation algorithm that utilizes the ratings or behavior of other users within the system to generate predictions and recommendations. The fundamental assumption behind this method is that different users’ opinions can be selected and aggregated in such a way as to provide a reasonable prediction of the active user’s preference [6].

The developers of the first RS in 1992 [7] coined the term collaborative filtering. The hypothesis of collaborative filtering is:

‘If user u and u’ rate n products, which are similar or have similar behavior, such as: viewing, buying, listening. Then they will have ratings for n products identical for other products’.

C. The k-nearest neighbor algorithm in collaborative filtering

The most common approach to CF is based on neighborhood models. There are two approaches
to collaborative filtering with the k-nearest-neighbor algorithm. First is user-based (User_kNN) prediction based on a similarity between users, and second is item-based (Item_kNN) prediction based on the similarity between items. User_kNN technique determines the similarity between two users by comparing their ratings on the same product. Then it predicts the rating on the product i by user u through the ratings of users similar to user u. The similarity between user u and user u’ can be calculated according to the equation Pearson (1) [6, 8, 9].

\[
\text{sim}_{\text{pearson}}(u, u') = \frac{\sum_{i \in I_{uu'}} (r_{ui} - \bar{r}_u)(r_{u'i} - \bar{r}_u)}{\sqrt{\sum_{i \in I_{uu'}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{uu'}} (r_{u'i} - \bar{r}_u)^2}} \tag{1}
\]

Where:
- \( r_{ui} \) and \( r_{u'i} \) are the ratings of users u and u’ on item i, respectively;
- \( I_{uu'} \) is the set of items rated by both user u and user u’;
- \( \bar{r}_u \) is the average rating value of all users with items u;
- \( \bar{r}_u' \) is the average rating value of all items with user u’.

With the User_kNN method, the rating review of user u on the product is predicted by Equation (2).

\[
\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{u \in Ku} \text{sim}(u, u')(r_{u'i} - \bar{r}_u)}{\sum_{u \in Ku} \text{sim}(u, u')} \tag{2}
\]

Where:
- \( \text{Sim}(u, u') \) is the similarity between users u and u’ determined by Equation (1);
- \( K_u \) is the number of users with proximity to user u (k neighbors of u).

D. System development and data collection

The research built a website for an English center with users who access the system and products as goods on the system. Users have attributes including User id, Username, Password, Full name, Avatar, Phone, Email, Coin, Role, and Active. Products have attributes including Product id, Product name, Product link, Product slug, Description, Thumbnail, Price, Price sales, Achieved, Unzip pass, View, Download, Active.

The website sells online courses, English self-study materials, and exam registration according to the competency framework. Users can view the services available on the web without registering as a member or registering as a member to get more benefits if they are non-members—the benefits like free documents, discounts, and especially recommended products on the service. Figure 2 shows the data model of website system.

E. Applying the k-nearest neighbor algorithm

The study applied the k-nearest neighbor algorithm to recommend documents on the web for users. There are three cases.

Case 1: If the users are not a member of the center, the system will recommend the most-read documents;

Case 2: If the users are members, the system will use the recommender system as described;

Case 3: If the users view many documents, the system will recommend unread or the latest document.

Figure 3 presents algorithm which is applied for three cases above.

III. RESULTS AND DISCUSSION

This article is applied according to the RS according to the collaborative filtering with kNN. Collaborative filtering is applied to analysis and
synthesis of the rating scores of the users, recognizes the similarity between users based on their rating scores, and generates recommendations based on this comparison. The study used a sample dataset of 129 users and 213 products with 1407 view data from the website.

There are two cases of the RS:

Case 1: The system will suggest the most favorite products when the user is not logged in. The most favorite documents have the most views from other users.

Case 2: When the user is logged in there are three clear cases:

- If the user is a newly registered account, their search history is unavailable. The RS can not be applied to the user so it will recommend the most favorite documents to them.

- If users view some documents on the website, the RS will identify their viewing history to recommend appropriate documents based on the RS according to collaborative filtering.

- If those who view all documents, it will be impossible not to find the data, so it will recommend documents that have never been viewed.

Suggestion Algorithm Processing

- Step 1: Get the search history of the user to recommend;

- Step 2: Compare the similarity of the current user’s viewing history with the viewing history of the other users in the system. If the similarity of 2 users is 50% or more, it is considered appropriate;

- Step 3: If the appropriate number of users is zero, go to step 5;

- Step 4: Retrieve all products according to the user’s viewing history with a similarity of 50% or more. Go to step 6;

- Step 5: Get all documents on the website except the user viewing history to recommend;

- Step 6: Return the documents to the user and finish.

Figure 4 shows the recommendation results for the case users who are non-members. Non-members visit the website but do not register as system members. In this case, the system will suggest the most viewed documents. Because this is a new user, they do not have any viewing history. So, there is no data to apply the algorithm. The result recommends the most viewed documents.

Suppose the users view the website and reg-
ister as a system member. Then the system will have data about the user. When users are system members, the system will have appropriate suggestions. Each time they visit the site, they will log into the system, the system determines the document viewing history, and the algorithm is applied to get new suggestions after each login. Figure 5 presents the recommendation results for the case users who are members using the RS.

Figure 5 shows that the most viewed documents are sorted in ascending order. That means highly regarded records with a statistically significant number of views have shown that these documents have a high viewership similarity to the user. That is consistent with the applied algorithm.

When a user is a system member and has already viewed most of the available documents, therefore, that becomes challenging for the system to identify other users with similar preferences. So, finding appropriate recommendations becomes difficult due to the algorithm calculating low similarity scores. Consequently, the system may suggest documents the user has never encountered. Figure 6 shows the recommendation results for the case in which the users are members who view the most of documents. The recommends the latest or unwatched documents.

Figure 6 shows suggested documents with low view. That means these documents are of little interest to other users and users who are not currently reading them.

Examining both scenarios of new users and existing members within the system reveals instances where document recommendations have been generated through the applied algorithm.

IV. CONCLUSION

The collaborative filtering model and product recommendation methods are potential research and practical application areas. The research will contribute to the development of practical solutions in this area.

A collaborative filtering method based on the kNN was applied based on a similarity between users or products. The RS can predict their rating with new documents by calculating user similarity. For new users, the RS uses attributes to recommend appropriate documents. In the future, the research direction is to test and re-evaluate the algorithm based on actual data from the online system after a period of operation and improve the algorithm. The study wants to explore other collaborative filtering methods and compare them...
with the neighbor model to develop the best method in different cases.

Moving forward, the study will focus on enhancing the algorithm by extracting user and document features to offer personalized recommendations. These improvements aim to improve the system’s convenience and provide users with a more favorable experience. By incorporating suggestions, the RS offers users more benefits than a scheme without personalized recommendations.

REFERENCES


