Intelligent Application of Digital Information Technology Based on Top-N Recommendation Method

Liyan Hu[✉]

Department of Economics, Shanghai University, Baoshan, Shanghai, China
Hh1354267@126.com

Abstract. With the rapid development of the Internet in China and the further improvement of the penetration rate of mobile phones, the base of mobile phone netizens is increasing day by day. The news client has appeared on the media stage as a new thing. With the sharp increase in the amount of data, traditional collaborative filtering algorithms have problems of sparsity and cold start. Based on this, this study proposes an improved item-based collaborative filtering model. Firstly, descriptive statistical analysis is carried out on the user's historical data to grasp the basic situation of the user's reading news information. After that, a user-item matrix based on implicit scoring is constructed. The item similarity is calculated by the cosine similarity formula, and the top 5 pieces of information with high interest are selected to recommend to users. Use precision and recall to evaluate recommendation performance. In addition, the mode method and Top-N recommendation method are used to solve the problem of user cold start. Finally, the follow-up research points need to be paid attention to, so as to promote the further development of news client recommendation.

Keywords: News Information · Intelligent Application · Top-N Recommendation Method

1 Introduction

With the advent of the Internet era, news clients have become an important means for people to obtain news information. In May 2022, the 49th "Statistical Report on Internet Development in China" released by China Internet Network Information Center (CNNC) showed that by December 2021. The number of Internet users in my country reached 1.032 billion, an increase of 42.96 million compared with December 2020. The Internet penetration rate reached 73.0%, and the proportion of Internet users using mobile phones to access the Internet reached 99.7%. The Internet penetration rate is increasing year by year, and the number of Internet users is increasing.

The rapid growth of smartphone users has led to a surge of various clients, such as Yahoo and Toutiao. The way netizens obtain news information has shifted from traditional media to terminals. The personalized recommendation algorithm is to generate different pages and recommend different content for different users by collecting data

such as user hobbies and usage habits. The intelligent application of digital information technology can place the content that everyone is most interested in and needed most on the recommendation page, so as to achieve accurate recommendation.

There have been more than 20 years of research on recommender systems, and a relatively complete theoretical foundation and framework has been formed, which is an indispensable part of network technology. However, recommender systems face the problems of cold start and sparsity. Aiming at the shortcomings of traditional collaborative filtering algorithms, this study analyzes the implicit meaning of user historical operation behavior data and mines its potential interest points to help users better obtain news information.

2   Literature Review

Collaborative filtering is the most famous recommendation algorithm in recommendation systems and has the most extensive applications. With the rapid growth of network data, the collaborative filtering recommendation algorithm has gradually exposed its application drawbacks, and scholars from all over the world have made their own attempts to optimize and expand it. For cold start problems, Liu and Lee (2009) applied the user's social network information to the recommender system, and found the nearest neighbors by measuring the user's similarity in the social information, so as to recommend items for new users [1]. Liang et al. (2020) used the optimized dynamic clustering algorithm to cluster user attribute characteristics, and formed a weighted similarity recommendation algorithm prediction model based on user attribute characteristics [2]. For data sparsity, Najafabadi et al. (2017) recommend users by mining the user's implicit information and the association between items, and propose to combine different rating matrices to fill the user's preference information [3]. Zhou et al. (2020) used the optimal boundary completion method and predicted score rounding to fill in missing values [4].

In recent years, recommender systems have been widely used in music, movies, news and many other fields [5–8]. The news client establishes the user's behavior database with the help of the recommendation algorithm, and makes personalized recommendations, so that the news client understands the user better. Dang Jun (2015) analyzed user characteristics and behaviors, and proposed differentiated push, emphasis on social experience and personalized customization for news clients [9, 10].

To sum up, in view of the problems existing in collaborative filtering algorithms, most scholars optimize the sparsity problem from the perspectives of dimensionality reduction, and improve the cold start problem from the perspectives of clustering and integrating social network information. However, these literatures don't comprehensively consider. At present, the research on news clients mainly analyzes user behavior satisfaction. In the field of algorithmic recommendation, there is little research in academia. This study uses the improved collaborative filtering recommendation algorithm to make accurate recommendation of news clients, which is of great supplementary significance.

3 Theoretical Analysis

3.1 Item-Based Collaborative Filtering Algorithm

The specific steps of the item-based collaborative filtering model are as follows:

(1) (1)Get the user-item rating matrix.
(2) (2)Use the cosine similarity formula to calculate the similarity of different items.

$$W_{jj} = \frac{|N(i) \cap N(j)|}{\sqrt{N(i)| \times |N(j)|}} \tag{1}$$

Among them, $|N(i)|$ and $|N(j)|$ are the number of users who have behaviors on item i and item j, respectively. $|N(i) \cap N(j)|$ is the number of users who have acted on items i and items j at the same time.

(3) Generate recommendation list. After the similarity is obtained, the user's interest in each item is sorted, and the top N item is selected and recommended to the user.

$$P_{uj} = \sum_{i \in N(u) \cap S(j,K)} w_{ji} r_{ui} \tag{2}$$

Among them, P_{uj} is user u's interest in item j, $N(u)$ is the set of items that user u has read, $S(j, K)$ is the set of K items most similar to item j, and r_{ui} is user u's rating on item i.

3.2 Improvement of Collaborative Filtering Algorithm

(1) User Implicit Behavior Score

The manifestations of user behaviors in recommender systems are divided into explicit and implicit behavioral data. The explicit behavior data is expressed as the user's direct score on the news information. The implicit behavior data is derived from the user's historical operation behavior data.

(2) Top-N Recommendation Method

There are two relatively simple ways to solve the problem of user cold start. One is the random recommendation method, that is, random recommendation is made to new users in all projects, and personalized recommendation is made according to the feedback of new users. The other method is Top-N recommendation. The method refers to recommending N most likely favorite contents to each user by mining the user's historical preference information [11, 12]. The hot spot in the information field is strong, and it is more effective to recommend hot information to new users. Therefore, this study adopts the Top-N recommendation method, that is, recommends the Top-N items in the popular list to new users.

3.3 Collaborative Filtering Algorithm Evaluation Index

This study selects the precision rate and the recall rate to measure the recommendation effect of the collaborative filtering algorithm. Suppose the recommendation system recommends N pieces of news information for each user, R_u is the recommended news information set based on the performance of user u, T_u is the news information set of user u 's actual behavior, and I represents the number of news information in the recommendation system.

The recall rate represents the percentage of useful recommendations in the news information that user u actually acts, and the calculation formula is shown in (3).

$$\text{Prcall} = \frac{\sum_u |R_u \cap T_u|}{\sum_u |T_u|} \tag{3}$$

The precision rate represents the percentage of useful recommendations in the recommended news information given according to user u's performance. The calculation formula is shown in (4).

$$\text{Precision} = \frac{\sum_u |R_u \cap T_u|}{\sum_u | R_u} \tag{4}$$

## 4 Data Sources and Descriptive Statistics

4.1 Data Sources

The research data comes from the data of the "Daguan Cup" big data algorithm competition, which has been released and shared on the official website. In order to protect the privacy of all users, the data has been desensitized.

4.2 Descriptive Statistical Analysis of Data

(1) Analysis of User Behavior Types

As shown in Fig. 1, among the types of user behaviors, "view" accounted for the largest proportion, at 55.49%, because not every piece of information read by users is of interest to users, and most users habitually read information roughly. Among the 327,043 user behaviors, "share" accounted for the least, only 0.02%. If news information is valuable and practical, that is, it meets the needs of users and is beneficial and useful to users, it will lead users to comment. High-quality content can not only increase user stickiness and make them stick around, but also stand out.
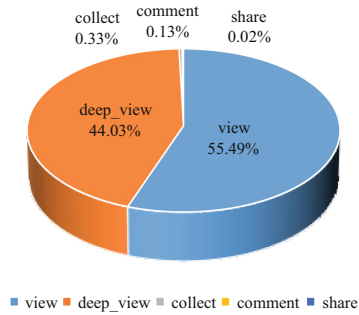
Fig. 1. Proportion of user behavior types

(2) User Activity Analysis

In the dataset Train, the total number of users is 28479, because the more users who read, the higher their activity. In this study, the six users who read the most news information are sorted. The results are shown in Table 1. It can be found that these users read more than 340 articles, and read more than 120 articles per day on average.

Table 1. Activity ranking top6

Rank	User_id	Reading volume	Rank	User_id	Reading volume
1	12944637	515	4	14699755	370
2	16026855	410	5	4023368	363
3	10005109	398	6	14684194	363

Define 05:00–12:00 as "morning", 12:00–18:00 as "afternoon", 18:00–22:00 as "evening", and 22:00–05:00 as "late at night". User_id 12944637 is a super active user of the news client. He is the most active in reading news and has read 515 articles in three days. Through analysis, it can be found that the user tends to read late at night, accounting for 40.22%. Recommending news to the user during this time period may increase the probability of successful recommendation (see Fig. 2).

(3) Analysis of Popular News

The reading volume of users is an important reference for various news clients to consider the "good" and "bad" of news information, and "good" refers more to the popularity. To roughly analyze the distribution of news readings, 5 news articles were randomly selected. The reading volume of popular news is very high, such as item_id 127350, up to 242. Conversely, item_id 112434 has only 1 reading volume. It may be that some news can significantly increase the reading volume by embedding hot words in the headlines.
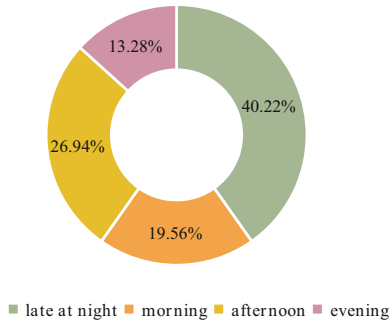
Fig. 2. The reading volume of news information of user ID 12944637 in different time periods

People are curious and have a deep understanding of many hot and secret events. On the other hand, it may be because some news itself is relatively unpopular, or the content of the news is not good enough to be read, and the reading volume of the news will be very small. A title with a hot word and a title without a hot word may affect the information by a difference of tens of thousands, hundreds of thousands or even hundreds of thousands of readings.

5 Intelligent Application of Digital Information Technology

5.1 News Client Recommendation Model Construction and Evaluation

(1) User Implicit Behavior Score

The presentation of news information is related to the content of interest, which will exacerbate the sparseness of data to a certain extent. In the dataset Train, a total of 65215 pieces of news information will form a matrix of 65215×65215, which is too much work and difficult to display. Therefore, to make the results accurate and reliable, users who read less than five articles were excluded from the original data. At the same time, 3000 users were randomly selected under the condition of guaranteed representativeness, and named dataset A. To quantify the user's implicit feedback behavior, assign "view = 1, deep_view = 2, collect = 3, commet = 4, share = 5", and the variable is named Rating. If the user has multiple behaviors, such as not only "view", the highest score will be calculated.

(2) Item-Based Collaborative Filtering Recommendation

Item similarity measurement is the core of the collaborative filtering algorithm. The Pearson correlation coefficient adopts the decentralized normalized dot product. Considering the uniformity of user scoring standards and the problem of sparsity, the cosine similarity calculation formula (1) is selected. Some of the similarity results are shown in Table 2.

Table 2. Similarity matrix results (part)

Item_id	Item_id	Similarity
437650	441211	0.5794
437650	446160	0.3403
437650	526888	0.2584

Eliminate the news information that the user has read in items to be recommended, calculate the user's interest in each piece of news information by formula (2), and recommend five pieces of news information for each user (see Table 3).

Table 3. Recommendation results and user interest level of user ID 329775

User_id	Action_type	Similar Item_id	similarity	degree of interest
329775	2	471862	0.8151	1.6303
329775	2	437320	0.7998	1.5997
329775	2	502659	0.6241	1.2483
329775	2	533295	0.6098	1.2196
329775	2	443834	0.6038	1.2076

(3) Evaluation of the Effect of Personalized Recommendation

The dataset Test gives all the news information of some users who have generated user behavior in N + 1 days, with a total of 1499 users. Take the common users in the dataset Test and A, a total of 237 users. By evaluating the results of the recommended news information, we can measure the effectiveness of the models and algorithms. According to the characteristics of the dataset and evaluation indicators used, two indicators, recall rate and precision rate, are selected. The precision rate of personalized recommendation is 22.53%, and the recall rate is 7.49%.

5.2 Recommend News to New Users

(1) Top-N Recommendation Method

Top-N is a commonly used means of information push, which is suitable for the situation where the recommendation system cannot obtain users' specific information. In the dataset, users have only one ID and no other information. Due to the strong hotness of the news field, it is more effective to recommend popular information to new users.

Therefore, the Top-N recommendation method is selected to recommend the top 5 news information in the hot search list to new users (see Table 4).

Table 4. The most popular news top5

Item_id	Cate_id
536585	1_17
536769	1_6
536551	1_1
536443	1_6
535553	1_6

(2) Mode method

Because most people have a mass mentality, the majority method is adopted based on this mentality of the people. Top-N recommendation has the disadvantage of a single recommendation category. In some cases, a certain category is particularly popular with users, which will lead to more popular news information of this type, and most of the news information recommended to users belongs to this type of news. Therefore, considering categories when recommending new users can further improve the probability of successful recommendation. This study selects the most popular piece of information in various categories.For example, from cate_id 1_6, the recommended item_id is 536769.

6   Research Conclusions and Prospects

This study proposes a comprehensive and improved method for the shortcomings of traditional collaborative filtering algorithms, to help users better obtain news information. Through the analysis, the following conclusions are drawn: (1)In view of the sparsity problem of traditional collaborative filtering algorithms, a user-item matrix based on implicit scoring is proposed to screen out the top 5 news articles with high interest and recommend them to users; (2)Aiming at the user's cold start problem, the majority method and the Top-N recommendation method are used to recommend new users. Top-N recommendation is suitable for the situation where the recommendation system cannot obtain user data, but it has the disadvantage of a single recommendation category. In some cases, a certain category is particularly popular with users, which will lead to more popular news of this type. Most of the user's news information belongs to this type of news, so the majority method is proposed. Considering the category when recommending new users can further improve the probability of successful recommendation.

However, there are still some shortcomings: (1)There is no attribute characteristic of the user involved. If user information can be collected, users with similar interests can be clustered to further recommend new users, which can further improve the cold

start problem; (2) There is no way to get more data. At present, in the field of news, many users will comment on the news information they are interested in. If the data of text comments can be extracted according to the content of the comments, combined with the implicit scores of users, the comprehensive implicit score data can be formed. Improve the existing model and make more accurate recommendations to users.

References

1. Liu, F.K., Lee, H.J.: Use of social network information to enhance collaborative filtering performance. Expert Syst. Appl. **37**(7), 4772–4778 (2009)
2. Liang, L.J., Li, Y.G., Zhang, N.N., Zhang, X., Wang, D.: Collaborative filtering algorithm integrating user features to optimize clustering. J. Intell. Syst. **15**(6), 1091–1096 (2020). (in Chinese)
3. Najafabadi, M.K., Mahrin, M.N., Chuprat, S., Sarkan, H.D.: Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on iimplicit data. Comput. Hum. Behav. **67**(2), 113–128 (2017)
4. Zhou, Y.T., Zhang, X.Y., Li, F.l., Liu, S.C., Jiao, J.l., & Tian, D.: SVDPP recommendation algorithm optimized by q-learning algorithm. Comput. Eng. **47**(2), 46–51 (2021). (in Chinese)
5. El, A.W.H., Ali, A.F., Slowik, A.: Hybrid crow search and uniform crossover algorithm-based clustering for top-N recommendation system. Neural Comput. Appl. **33**(12), 7145–7164 (2020)
6. Ethem, C., Ibrahim, M., Tevfik, A.: Integrating Individual and aggregate diversity in top-n recommendation. INFORMS J. Comput. **33**(1), 300–318 (2020)
7. Daniel, V., Alejandro, B., Javier, P., Pablo, C.: Assessing ranking metrics in Top-N recommendation. Inf. Retrieval J. **23**(4), 1–38 (2020)
8. Sharon, M.J., Dhinesh, B.L.D.: Genetic algorithm influenced Top-N recommender system to alleviate new user cold start problem. Int. J. Swarm Intell. Res. **11**(2), 62–79 (2020)
9. Dang, J.: User experience analysis of mobile news client. Friends of Edit **14**(9), 62–64 (2015). (in Chinese)
10. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. Res. Dev. Inf. Retrieval **51**(2), 230–237 (1999)
11. Jin, R., Si, L., Zhai, C.X.: A study of mixture models for collaborative filtering. Inf. Retrieval **9**(3), 357–382 (2006)
12. Koren, Y., Bell, R.M., Volinsky, C.: Matrix factorization techniques for recommender systems. IEEE Comput. **42**(8), 30–37 (2009)