Collaborative filtering recommendation algorithm based on user behavior and drug semantics

Ruiyang Wang, Tao Li*

College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China

*Corresponding author: litaow@wust.edu.cn

Abstract. When the traditional collaborative filtering algorithm is applied to drug recommendation, the recommendation effect is not good due to the sparsity of data. In view of the above problems, this paper proposes a collaborative filtering recommendation algorithm based on user behavior and drug semantics (UBDS-CF). Firstly, we construct the purchasing behavior matrix of users and drugs, and use the weighted cosine similarity to calculate the basic similarity between drugs; then construct the category label matrix of drugs, calculate the category similarity, and extract the feature vector of drug function text using the word vector model; category and main function together constitute the drug semantic information; finally, the above two similarities are integrated in a linear weighted manner to obtain the final similarity of drugs. Experimental results show that the average absolute error and root mean square error of the algorithm are significantly reduced, which proves the effectiveness of the proposed algorithm.

1. Introduction

In recent years, with the rapid development of medical level, the types and quantity of drugs are increasing day by day. It is very difficult for users who buy medicine online to choose the most effective drug for their own diseases from the complex drugs. Personalized recommendation system [1] is a technology of filtering information resources and pushing data correctly. We apply the recommendation algorithm in the field of drugs, which can help users to buy the drugs they need.

Personalized recommendation algorithms mainly include: content-based recommendation algorithm, collaborative filtering based recommendation algorithm and hybrid recommendation algorithm. Collaborative filtering [2] is not limited by professional knowledge, and can use a variety of unstructured data, so it is widely used in e-commerce, social media, video music and other platforms. However, the traditional collaborative filtering recommendation algorithm cannot solve the problem of data sparsity, which leads to the low accuracy of recommendation.

To solve this problem, scholars at home and abroad have carried out in-depth research from different angles, among which, matrix filling and user (commodity) clustering method are the research focus of the majority of scholars.

Matrix filling is a method of using a certain strategy or algorithm to get the estimated value of missing data to replace the missing value. Reference[3] takes into account the impact of sparsity on the algorithm, uses fixed values such as the average rating of users or items to fill in the score matrix, and accurately
calculates the similarity; Reference[4] fills in by mining hidden information in Web logs The matrix then
obtains a complete scoring matrix, which effectively solves the problem of a large amount of missing
data in the matrix; in order to prevent the use of a fixed value filling matrix to introduce new noise,
Reference[5] proposes an improved weighted naive Bayes algorithm to predict unrated Data method;
Reference[6] uses the Slope One algorithm to predict missing values and fills the user rating matrix;
finally, a hybrid collaborative filtering algorithm is used to perform a nearest neighbor search on the filled
user rating matrix.

Some scholars perform clustering based on the characteristic attributes of users or items, and obtain
sub-matrices that are denser than the original score matrix to alleviate the sparsity problem. Reference[7]
applies dimensionality reduction and clustering technology to collaborative filtering recommendation
algorithm, uses mean algorithm and singular value decomposition to cluster and dimensionality reduction
of similar users, and obtains accurate and efficient recommendation results; Reference[8] The rating
establishes a corresponding incentive model to obtain different preferences among users, clusters the
different preferences of users, and then recommends among user groups with the same preference;
Reference[9] believes that similar products are more similar, based on item category attributes. Clustering,
recommending items by category to users.

Although both matrix filling and clustering methods can effectively alleviate the impact of data
sparsity, matrix filling does not take into account the differences of users with different preferences, and
it is easy to introduce new noise; clustering is similar to down sampling, which will lose some important
sample. Therefore, this paper proposes a collaborative filtering algorithm based on user behavior and
drug semantics. First, construct the purchase behavior matrix of users and drugs to calculate the initial
similarity of drugs; then construct the drug-category label matrix, and use the word vector model to
extract the text feature vectors of the drug’s main functions to calculate their similarities; finally, use
linear weighting The two similarities are integrated together to correct the error that exists when using a
sparse matrix for recommendation.

2. Related work
The collaborative filtering algorithm[10] is mainly divided into three steps: firstly, use the user’s
interaction data (such as: click, browse, purchase, rating) to construct a user-item matrix, and then use
the similarity calculation method to find the nearest neighbor set of the user or item according to the
constructed matrix, and finally get the neighbor’s feedback data. The association between the unknown
user and the item, resulting in a recommendation list. The calculation of similarity is the key to the
collaborative filtering algorithm, and the quality of the similarity calculation will directly affect the
quality of recommendations. Currently commonly used similarity calculation methods are: cosine
similarity, Jaccard similarity and other methods [11]:

Cosine similarity regards the user's rating sequence of items as vectors, and the cosine value of the
angle between two vectors can be used to indicate the similarity of different users or items. The smaller
the angle, the closer the cosine value to 1, and the higher the similarity. The cosine similarity calculation
formula is as follows:

\[
\text{sim}(I_i, I_j) = \frac{\sum_{k=1}^{n} r_{I_i,U_k} r_{I_j,U_k}}{\sqrt{\sum_{k=1}^{n} r_{I_i,U_k}^2} \sqrt{\sum_{k=1}^{n} r_{I_j,U_k}^2}}
\] (1)

Where \(r_{I_i,U_k}\) represents the rating value of user \(k\) on item \(i\).

The Jaccard coefficient is the ratio of the intersection and union of two scoring sets, which is suitable
for calculating the similarity of discrete scoring data. The calculation formula of Jaccard coefficient is as
follows:
3. Collaborative filtering algorithm based on user behavior and drug semantics

3.1. Similarity based on user behavior

This article constructs a user-drug purchase behavior matrix based on the sales of drugs. 0 represents that the user $U_k$ has no purchase behavior for the drug $I_i$, and 1 represents that the user $U_k$ has historical purchase behavior for the drug $I_i$. Then we need to choose an appropriate similarity calculation method to calculate the basic similarity between drugs. Euclidean distance is suitable for calculating the distance between the attribute features of two items to measure the similarity of the two, and the Person similarity calculation method will decentralize the data in the calculation process, and is more sensitive to the explicit score data [12]. Therefore, this article uses the cosine similarity calculation method when calculating the similarity of drugs based on the sales of drugs.

This article believes that whether a user purchases a drug reflects the user's interest and preference, and the frequency of user purchases of this drug reflects the user's preference to a certain extent. Interest preference and degree of preference work together to reflect the similarity in sales behavior of the two drugs. When users have a demand for a certain medicine, they will inevitably repeat purchases within a certain period, and there are obvious characteristics of purchase frequency. Therefore, when using the cosine similarity coefficient to calculate the similarity of medicines, this article uses the purchase frequency as a weighting factor to improve the accuracy of the similarity calculation. The formula is as follows:

$$\text{sim}(I_i, I_j)_{\text{fre}} = \frac{\sum_{k=1}^{n} e^{-|r_{I_i, U_k} - r_{I_j, U_k}|} r_{I_i, U_k} r_{I_j, U_k}}{\sqrt{\sum_{k=1}^{n} r_{I_i, U_k}^2} \sqrt{\sum_{k=1}^{n} r_{I_j, U_k}^2}}$$

(3)

Among them, $r_{I_i, U_k}$ indicates whether user $U_k$ has a purchase behavior for drug $I_i$, $r_{I_j, U_k}$ indicates the purchase frequency of user $U_k$ for drug $I_j$. It can be seen that the greater the difference between the purchase frequency of two drugs When the time, the similarity between the two is smaller. The new similarity calculation formula takes into account the impact of user purchase behavior on drug similarity, making the calculation of drug similarity more in line with real-life scenarios.

3.2. Similarity based on drug semantics

This article improves the cosine similarity by taking the purchase frequency as the weight, which improves the accuracy of the calculation of the medicine similarity to a certain extent. However, the drug-user purchasing behavior matrix is relatively sparse, and using only a small amount of data for predictive recommendation cannot guarantee the quality of the recommendation system. Reference [13] pointed out that comprehensive consideration of user dynamic behavior data and item static attribute characteristics can alleviate the problem of inaccurate similarity measurement caused by data sparseness to a certain extent. In the case of less user behavior data, item attributes can be used to calculate the similarity of item attributes to reduce the error of similarity calculation. By digging out the static properties of drugs and comparing them with other products, this article finds that drugs have clear classification properties and efficacy, and drugs with the same effect have extremely high similarities in knowledge.

Assuming that the classification set of medicines is \(C = \{C_1, C_2, C_3 \ldots C_n\}\), where \(n\) represents the number of classifications, a medicine classification matrix \(R_{m \times n}\) can be constructed, and \(m\) is the number of medicines. The formula for calculating the category similarity of drugs is as follows:
\[
\text{sim}(I_i, I_j)_{\text{cla}} = \frac{|C_i \cap C_j|}{\max(|C_i|, |C_j|)}
\]

(4)

Where \( C_i \) is the classification set of the drug \( I_i \), and \( |C_i| \) represents the number of classifications of the drug \( I_i \). Formula (4) is different from the Jaccard similarity coefficient, because there are many drug categories, it is difficult for two drugs to have a high consistency in the category, and the use of intersection as the denominator will make the calculation result of the similarity of the drug category smaller.

The main function of drugs is stored in the text. In order to obtain the similarity of drug efficacy, this article uses the third-party library of Chinese word segmentation jieba to segment the main functions of drugs. After removing the stop words, the phrase list is obtained. Use word2vec to train the word vector for the obtained phrase list, and obtain a K-dimensional word vector. Each dimension in the vector corresponds to a feature word in the text description of the medicine’s indication function. Then use the Simhash algorithm [14] to calculate the similarity, the formula is as follows:

\[
\text{sim}(v_i, v_j)_{\text{fun}} = \sum_{k=1}^{K} \frac{v_{i,k} \oplus v_{j,k}}{K}
\]

(5)

where, \( v_i, v_j \) represents the simhash value of drug \( I_i \) and drug \( I_j \), and \( v_{i,k}, v_{j,k} \) represents the k-dimensional value of each vector.

So far, the similarity of the same drug can be expressed as:

\[
\text{sim}(I_i, I_j)_{\text{cla.fun}} = \frac{|C_i \cap C_j|}{\max(|C_i|, |C_j|)} + \sum_{k=1}^{K} \frac{v_{i,k} \oplus v_{j,k}}{K}
\]

(6)

3.3. Comprehensive similarity of drugs

The basic similarity of drugs is the degree of similarity of drugs extracted from the user’s behavior in purchasing drugs, while the semantic similarity of drugs reflects the degree of similarity of drugs in categories and functions. The single use of any of the above similarities cannot be compared. Good reflects the degree of similarity between the drugs. Therefore, this paper weights the two similarities linearly to obtain the final similarity of the drugs. The calculation formula can be expressed as:

\[
\text{sim}(I_i, I_j) = \alpha * \text{sim}(I_i, I_j)_{\text{fre}} + (1 - \alpha) * \text{sim}(I_i, I_j)_{\text{cla.fun}}
\]

(7)

\[
\text{sim}(I_i, I_j) = \alpha \sum_{k=1}^{n} e^{-|t_{i_k} - t_{j_k}| r_{i_k} - r_{j_k}} + (1 - \alpha) * \frac{|C_i \cap C_j|}{\max(|C_i|, |C_j|)} + \sum_{k=1}^{K} \frac{v_{i,k} \oplus v_{j,k}}{K}
\]

(8)

Where \( \alpha \) is the weight factor, the value range is \([0,1]\), the specific value is obtained by experiment. The algorithm in this article can be summarized as follows:
Table 1. Algorithm flow

| Algorithm: Collaborative filtering recommendation algorithm based on user behavior and medicine semantics (UBDS-CF) |
|---|
| **input**: drug historical sales records, users, and drug attributes;  
**output**: recommended list of drugs; |
| **step 1** | A user-drug purchase behavior matrix R is constructed according to the sales records of the drugs. For each drug in the drug set, the basic similarity between drugs is calculated using formula (3). |
| **step 2** | Construct a classification matrix based on the drug classification data, and at the same time segment the text information of the drug function, and then use the word vector model to train to obtain the word vector of the drug. For each drug in the drug set, use the formula (6) to calculate the semantics between the drugs. |
| **step 3** | Normalize the above two similarities so that their values are mapped in the range of [0,1], and then linearly weighted to obtain the final similarity of the medicine. |
| **step 4** | For each medicine in the medicine set, rank the similarity between it and other medicines, select the k medicines with the largest similarity as neighbors, and obtain the neighbor set of each medicine. |
| **step 5** | Through the neighbor set of each medicine, the score of each medicine can be calculated by the target user. Repeat the above steps to get the target user's score prediction value for unrated drugs, and rank them to generate Top-N recommendations. |

4. Experiment and analysis

4.1. data set

The experiment in this article uses an online real data set provided by a chain pharmacy. The data set is the sales of main drugs in one store of the chain pharmacy, including drug information, member information and drug sales records; the size of the data is: 325 main drugs, 26937 members, and 94817 consumption records. The data sparsity is 98.9%. The experiment is conducted by randomly dividing 80% of the data as the training set and 20% of the data as the test set.

4.2. Evaluating indicator

In this paper, two indicators, MAE (Mean Absolute Error) and Root Mean Squared Error (RMSE) are used as the evaluation indicators of the algorithm. The average absolute error reflects the average value of the absolute difference between the predicted value of the algorithm and the user's real score value. The smaller the value, the more accurate the prediction. The root mean square error reflects the deviation of the predicted value from the true value. Their corresponding formulas are as follows:

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |p_i - r_i| \quad (9)
\]

\[
\text{RMSE} = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (p_i - r_i)^2} \quad (10)
\]

In formula (9) and formula (10), n represents the sample space size of the test data, \(p_i\) represents the predicted score generated by the algorithm, and \(r_i\) represents the actual score of the drug produced by the user.

4.3. Selection of \(\alpha\)

This paper constructs the purchase behavior matrix between users and drugs, and uses weighted cosine similarity to calculate the basic similarity between drugs; at the same time, considering the category and
functional characteristics of the drugs themselves, this paper uses the word vector model to extract text features between drugs, so as to calculate the semantic similarity between medicines. In order to measure the above two similarities uniformly, this paper uses $\alpha$ as a weighting factor to linearly weight the two. By adjusting the value of $\alpha$ and comparing the corresponding MAE and RMSE, the linear combination with the best experimental effect can be obtained.

It can be seen from Figure 1 that for different values of $\alpha$, the MAE and RMSE values are different, and the fluctuation range is relatively large, and the overall trend is to first decline and then rise. This shows that MAE and RMSE are very sensitive to the alpha value. When $\alpha=0.2$, MAE and RMSE get the minimum value, and the weight of the semantic knowledge of medicine accounts for 80%. It can be seen that the semantic knowledge of medicine can effectively reduce the error of the recommendation result. However, when $\alpha<0.2$, the weight of the similarity of the drug score is too small, and the recommendation is generated by relying on the semantic similarity of the drug too much, and the actual sales situation is not taken into account, resulting in very unstable recommendation results.

![Figure 1. The effect of $\alpha$ on error](image)

4.4. Selection of nearest neighbor number $k$

The number of neighbors $k$ represents the number of $k$ drugs that are most similar to a certain drug. When the number of neighbors $k$ is between 0-50, the performance of the algorithm in this paper on MAE and RMSE is shown in Figure 2.

It can be seen from Figure 2 that when the number of neighbors $k$ continues to increase, the effective data that this algorithm can use also increases, and the values of MAE and RMSE are gradually decreasing. When the number of neighbors exceeds 30, although the magnitude of the two error values fluctuates, they tend to be stable overall. Therefore, the value of the nearest neighbor number $k$ selected by the algorithm in this paper is 30.
4.5. Experimental comparison of different algorithms

In order to more accurately evaluate the quality of this algorithm (UBDS-CF), this paper designs several sets of comparative experiments. The methods compared in this paper are: traditional collaborative filtering recommendation algorithm (Item-CF), collaborative filtering recommendation algorithm based on mixed filling (HF-CF) [15] and a collaborative filtering recommendation algorithm based on user trust clustering (TC-CF) [16]. The experimental parameters in this article are set as follows: the weight $\alpha$ is 0.2, the number of neighbors $k$ is 30, and the recommended list length is 60. The experimental results are shown in Figure 3:

**Figure 3.** Comparison of error values of different algorithms

**Figure 2.** The influence of the number of neighbors $k$ on the error
It can be seen from Figure 3 that under the same experimental conditions, the two error values of the algorithm in this paper are lower than the other three algorithms. Compared with the traditional collaborative filtering algorithm, the MAE value of the algorithm in this paper is reduced by 12%, and the RMSE The value of is reduced by 10%; compared with the collaborative filtering hybrid filling algorithm, the RMSE and MAE values of this algorithm are reduced by 4% and 6%, respectively; compared with the collaborative filtering recommendation algorithm based on user trust clustering, the RMSE and MAE values of the algorithm are reduced by 3% and 5% respectively. The above experimental results show that the algorithm in this paper can effectively improve the quality of drug recommendations.

5. Conclusion
When recommending drugs, due to the large sparseness of the data in the real scene, the use of traditional collaborative filtering algorithms is prone to poor recommendation effects. In response to this problem, this paper proposes a collaborative filtering recommendation algorithm based on user behavior and drug semantics. The user-drug purchase behavior matrix is constructed based on drug sales, and the weighted cosine similarity is used to calculate the basic similarity between drugs, and then A drug-category label matrix is constructed, and the text feature of the drug function is extracted using the word vector model to calculate the semantic similarity between the drugs. Finally, the two similarities are linearly weighted to obtain the final drug similarity. Experimental results show that the proposed algorithm can effectively improve the accuracy of drug recommendation. However, the algorithm in this paper is still an improved algorithm for collaborative filtering. In recent years, methods based on matrix factorization and neural networks have also been widely used in recommendation systems. Therefore, in the future, this paper will continue to conduct in-depth research in this direction.

References
[1] Wang Z, Wan M, Cui X, et al. Personalized Recommendation Algorithm Based on Product Reviews[J]. Journal of Electronic Commerce in Organizations, 2018, 16(3):22-38.
[2] Shah K, Salunke A, Dongare S. Recommender systems: An overview of different approaches to recommendations[C]. 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). IEEE, 2018.
[3] Kant S, Mahara T. Merging user and item based collaborative filtering to alleviate data sparsity[J]. International Journal of System Assurance Engineering & Management, 2018, 9(1):1-7.
[4] Fuliang X. Research of collaborative filtering recommender method using WUM and RBFN to fill missing values[J]. Computer Engineering & Applications, 2012.
[5] Daxue Li. Collaborative filtering recommendation algorithm based on Naive Bayesian method[J]. Computer application, 2010, 30(06):1523-1526.
[6] Ming Geng, Zhenrong Deng. Collaborative recommendation algorithm based on user clustering and slope one filling[J]. Computer engineering and Application, 2018, 54(22):144-148.
[7] Hafed Zarzour, Ziad Al-Sharif, Mahmoud Al-Ayyoub. A new collaborative filtering recommendation algorithm based on dimensionality reduction and clustering techniques[C]//2018 9th International Conference on Information and Communication Systems (ICICS). IEEE, 2018.
[8] Tran C, Kim J Y, Shin W Y. Clustering-Based Collaborative Filtering Using an Incentivized/Penal- ized User Model[J]. IEEE Access, 2019:1-1.
[9] Yong Wang, Xiaoyi Wan, Yaozhi Tao. Collaborative filtering recommendation algorithm based on k-medoids item clustering[J]. Journal of Chongqing University of natural sciences and Telecommunications, 2017, 029(004):521-526.
[10] Karabadji, Nour. El. Improving memory-based user collaborative filtering with evolutionary multi-objective optimization[J]. Expert Systems with Application, 2018.
[11] Wenqiang Xiao, Shijun Yao, Shanming Wu. An improved top-N collaborative filtering recommendation algorithm[J]. Research on computer application, 2018, 35(01):111-114+118.
[12] Yuhang Xiao, Mingli Wu. Application of collaborative filtering algorithm to improve user similarity[J]. Information Technology, 2018, v.42; No.320(07): 138-142.

[13] JiaMing Wang, Jingchou Han. Collaborative filtering algorithm based on user preference for item attributes[J]. Computer Engineering and Application, 2017, 53(006): 106-110.

[14] Tiannan Wang, Feng Feng. Research on text similarity detection algorithm based on simhash[J]. Electronic testing, 2019(15).

[15] YongGong Ren, Siyu Wang, Zhipeng Zhang. Collaborative Filtering Hybrid filling algorithm for alleviating data sparsity[J]. Pattern recognition and artificial intelligence, 2020(2): 166-175.

[16] Fei Wang, Gang Huang, Zhenyu Zhu. Collaborative filtering recommendation algorithm based on trust clustering [J]. Computer technology and development, 2019, 29(05): 28-32.