ISSN: 2664-9640

Research on similarity algorithm of collaborative filtering for sparse data

Wentao Zhao, Tingting Feng, Ziheng Cui

College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo Henan 454000, China

Abstract

In the neighborhood-based collaborative filtering algorithm, similarity model plays a decisive role in the algorithm's recommendation performance. Traditional similarity model focuses on co-rated items, and the aggravation of data sparsity tends to reduce the accuracy of traditional similarity model and affect the reliability of nearest neighbor recommendation. In order to solve the problem of low reliability of neighbor recommendation in sparse data, firstly, Jensen-Shannon (JS) divergence is introduced as the basis function, and the global rating probability distribution is used to measure the preference similarity between users to alleviate the problem of data sparsity. Secondly, Structural similarity contains rating values was defining as a weighting factor that emphasize the importance of co-rated items. At the same time, a differentiation similarity calculations is designed for the co-rated items to improve the differentiation degree of similar users, obtain the similarity based on the relative interval span. Experiments on different sparsity datasets show that the proposed algorithm has better performance in both prediction and recommendation accuracy indicators than other algorithms

Keywords

Sparse data; collaborative filtering; js divergence; global structure; the range of span.

1. Introduction

The amount of information on the Internet increases exponentially, and users cannot quickly obtain interested information in massive data[1], resulting in information overload[2~4]. The recommendation system provides personalized services[5] for users based on the historical behavior data of users, filters redundant information from massive information, and recommends information that meets the needs of users to solve the problem of information overload. Recommendation algorithm[6,7] is the core element of recommendation system and determines the type and performance of recommendation system. Among them, neighbor-based collaborative filtering is a widely used algorithm in recommendation systems, and determining the nearest neighbor of target users based on similarity is a key part of it[8]. Therefore, similarity model plays a decisive role in prediction and recommendation results.

Traditional similarity models, such as Pearson Correlation Coefficient (PCC)[9], Cosine Correlation Coefficient (COS)[10] and Adjust Cosine, ACOS][11] and Jaccard coefficient[12] (Jaccard), etc, mined the contribution of common rating information, but lost the value of non-common rating information, resulting in low accuracy of prediction and recommendation under sparse data. In order to improve the accuracy of prediction and recommendation, many researchers have proposed new or improved similarity models based on traditional models to alleviate data sparsity and solve the cold start problem. For example, Weighted Pearson Correlation Coefficient (WPCC)[13], which is improved based on PCC[9], has a relatively higher recommendation accuracy.

ISSN: 2664-9640

The new linear heuristic similarity[14](PIP) shows a better recommendation performance in a cold start environment. But neither is free from the restrictions of the common rating program. New Heuristic Similarity Model[15] NHSM, BCF[16] (Bhattacharyya Coefficient) and KLCF[17] (KL-based Similarity Measure) all use global rating information. These new algorithms effectively alleviate the impact of data sparsity, but ignore the value of users without common rating items, and have high time complexity, which affects the accuracy of recommendation.

Based on the above analysis, a similarity model for sparse data is proposed. Firstly, from the perspective of rating probability distribution, a rating preference similarity based on improved JS divergence is proposed, and the rating quantity information is integrated to make up for the order of magnitude error, so as to make full use of all rating information and alleviate the impact caused by data sparsity. Secondly, the same rating set is defined, and the structural similarity of fused rating values is proposed as the weight factor, emphasizing the importance of common rating items, and improving the integrity of structural similarity and the accuracy of recommendation. User rating projects together in the end, be divided into the same subgroup and the same interval subgroup, and consider the similarity should satisfy the qualitative conditions, according to different design differentiation subgroup of similarity calculation methods, improve similar user similarity and the degree of differentiation, for similarity based on relative interval span, further enhance the reliability of the model.

2. Related work

Determining the target user's neighbor set is a key factor for the high performance of recommendation algorithm. In the collaborative filtering algorithm based on neighborhood, similarity model and data sparsity are important factors that affect the quality of recommendation. Therefore, improving the prediction and recommendation accuracy of recommendation algorithms by improving the similarity model[11] has become a research hotspot for researchers.

At present, researchers have provided a rich theoretical basis for collaborative filtering recommendation algorithms. Sun et al.[18] proposed a Triangle multiplying Jaccard (TMJ) similarity model combining Triangle and Jaccard. In this model, the Triangle takes the length and Angle of the project rating vector into consideration. The traditional similarity model breaks the restriction that only considers Angle[10] or length[19] between rating vectors. However, the similarity model does not get rid of the dependence on common rating items. In order to break the limitation of common rating items, Liu et al.[15] proposed a new heuristic similarity degree NHSM, which is composed of PIP-standardized PSS[15] and user preference similarity expressed by URP, and has a good recommendation performance on cold start users. Jesus[20] et al. proposed a singularity point-based similarity model SM (SM), which calculated the singular values of users according to the singular points of each project, and verified the effectiveness of the model. KLCF, WAJS[21] and BCF algorithm[16] all measure the similarity between projects from the perspective of probability density distribution of project ratings, make full use of all ratings and improve the quality of neighbor recommendation, but KLCF and BCF have high computational complexity. Khrouf et al.[22] proposed a collaborative filtering recommendation algorithm for structured information, which omitted the process of information retrieval and introduced other data sources of social network information to make up for data sparsity, enrich data structure forms and improve the recommendation quality of neighboring neighbors. However, the accessibility of social network information affects the applicability of the algorithm.

Aiming at the problem of data sparsity, in order to make full use of all ratings and further improve the accuracy of the algorithm, this paper conducts further research on the recommendation algorithm based on neighborhood.

ISSN: 2664-9640

3. The proposed method

The proposed similarity model (JSR) consists of two parts. The first part is the rating preference similarity based on the improved JS divergence[23], and the second part is composed of the structural similarity based on the fusion of rating values (SJaccard) and the similarity based on the relative interval span. The final formula of the similarity model is shown in (1). Where is the parameter that adjusts the proportion according to different data sets.

$$sim(u,v)^{JSR} = \lambda \cdot sim(u,v)^{JS} + (1\text{-}\lambda) \cdot sim(u,v)^{SJ} \cdot sim(u,v)^{RIS} \tag{1}$$

3.1. Rating preference similarity based on JS divergence

In order to alleviate the problem of data sparsity and improve the reliability of similar users, the user rating probability and moderating factor are defined first, and then the rating preference similarity based on improved JS divergence is proposed from the perspective of rating probability distribution. JS divergence is used to measure the symmetry distance between two probability sequences.

Definition 1 (The probability of rating) Set of rating values $C = \{c_1, c_2, \ldots, c_i, c_{i+1}\}$. Users of u's non-zero rating set is $I_u = \{r_{u,i} \| i = 1,2,\ldots,n, 且 r_{u,i} \neq 0\}$. The proportion of each rating value in the set of rating values is the rating probability, Is defined as $\mu_u(c_i)$. The average density of rating values c_i is defined as $\omega(c_i)$, As shown in equations (2) and (3) respectively;

$$\mu_u(c_i) = \frac{|c_i^u|}{|I_u|} \tag{2}$$

$$\omega(c_i) = \frac{1}{2} \cdot \left(\frac{|c_i^u|}{|I_u|} + \frac{|c_i^v|}{|I_v|} \right) \tag{3}$$

$|c_i^u|$ represents the number of rating values among all non-zero rating values of user u, $|I_u|$ Represents the total number of rated items of user u.

The rating distance based on JS divergence is shown in Equation (4), which makes full use of all the rating information, which is conducive to alleviate the problem of data sparsity and improve the recommendation accuracy and prediction accuracy of the algorithm.

$$JS(u \| v) = \frac{1}{2} \cdot \left(\sum_{c_i \in C} \mu_u(c_i) \cdot \lg \frac{\mu_u(c_i)}{\omega(c_i)} + \sum_{c_i \in C} \mu_v(c_i) \cdot \lg \frac{\mu_v(c_i)}{\omega(c_i)} \right) \tag{4}$$

The rating preference distance proposed based on JS divergence indicates that the smaller the difference between the two probability sequences is, the higher the rating preference similarity between users is, that is, inversely proportional to the rating preference difference similarity of users. That is, the similarity of user rating preference is shown in Equation (5).

$$sim(u,v)^{JS} = 1 - JS(u \| v) \tag{5}$$

3.2. Structural similarity of fusion rating value SJaccard

The traditional structural similarity model (Jaccard coefficient) only considers the number of common rating items among users, and ignores the influence of the number of user rating values on the model. On the basis of Jaccard coefficient, when the rating values of the two items are equal, their contribution degree is consistent with that of the common rating items. In fact, the equal rating value indicates that the two users' preferences for this item are completely consistent, which is more convincing to enhance the similarity between users. Considering the

ISSN: 2664-9640

same number of rating values and the number of users' common rating items, a new structural similarity SJaccard is proposed.

Table 1 User-item rating matrix

	I1	I2	I3	I4	I5
u1	5	2	/	/	/
u2	5	3	/	/	/
u3	5	3	5	5	5
u4	5	5	/	/	/
u5	5	4	4	1	2

For further explanation, the rating matrix of five items by five users is selected, as shown in Table 1. The similarity between users in Table 1 is measured based on the Jaccard coefficient. The similarity between user u1 and user u3 is the same as that between user u2 and user u3,is $sim(u_1,u_3)^{Jaccard} = sim(u_2,u_3)^{Jaccard} = 0.4$. However, user u2 and u3 have the same rating on item I2, indicating that user u2 and u3 have the same interest in item I2, then user u2 and u3 have higher similarity. So, $sim(u_1,u_3) < sim(u_2,u_3)$. Based on the above analysis, in order to improve the recommendation performance of structural similarity, the same rating set is defined by comprehensively considering the number of items with the same rating value and the number of items with common rating value, and the structural similarity SJaccard fused with rating value is proposed.

Definition 3 (Same rating set) A set of items that are rated the same by different users is called the set of the same rated items, is $I_{u\|v} = \{i \| r_{u,i} = r_{v,i}, 且 r_{u,i} \neq r_{v,i} \neq 0\}$.

The value interval of the proposed new structural similarity SJaccard is, and the formula is shown in (6), where is the number of items rated the same by users u and v.

$$sim(u,v)^{SJ} = \frac{|I_u \cap I_v| + |I_{u\|v}|}{|I_u| + |I_v|} \tag{6}$$

According to Equation (6), similar users of U3 can be effectively distinguished, is $sim(u_1,u_3)^{SJ} \approx 0.429 < sim(u_2,u_3)^{SJ} \approx 0.517$. SJaccard not only considers the impact of the number of common rated items on the similarity, but also reflects the importance of the information of the number of items with the same rating. It enhances the reliability of the similarity of data structure, improves the recommendation accuracy of neighbor set, and reduces the prediction error of project rating.

3.3. Similarity based on relative interval span

In terms of rating probability, user rating preference based on improved JS divergence breaks the constraint of data sparsity, and effectively uses all rating information to accurately measure the difference in user rating preference. From the aspect of rating structure, SJaccard model eliminates the problem of unreliable similarity calculation caused by the unbalanced number of ratings. In order to further improve the prediction and recommendation accuracy of the model, and reflect the importance of the common rating items to measure the similarity, the similarity based on the relative interval span is proposed from the aspect of the rating value difference of the common rating items.

A rating is a numerical measure of how much a user likes an item, if the rating interval is $[1,5]$, the rating interval $[1,2]$ represents the user's negative preference interval for the project, the interval $[4,5]$ represents the user's positive preference interval, and the score 3 represents the

ISSN: 2664-9640

medium attitude. When different users' ratings of the same item are distributed in two different ranges, it indicates that users have different interests. In Table 1, the SJaccard similarity between u_1 and u_5, u_4 and u_3 is the same. And the difference in rating values relative to I_2 is also equal,is $|r_{1,2}-r_{5,2}|=|r_{4,2}-r_{3,2}|=2$. However, u_3 and u_4 rated item I_2 in different ranges, while u_1 and u_5 rated item I_2 in the same range, indicating that u_1 and u_5 were more similar. Considering the interval span influence on similarity measure, the user common rating program is divided into the same subgroup and the same interval subgroup, defining weighted similarity factor θ, when comparing similarity to the same interval of subgroup rating to weighted difference downgrade difference, similar to enhance user similarity and the degree of differentiation, promote neighbor recommendation quality.

The absolute value of rating difference is directly used to measure the similarity, and its similarity interval is [0,+∞], which will infinitely reduce the contribution of SJaccard and JS similarity model to the similarity, which is not practical. In order to standardize the similarity of the three models within a unified interval and satisfy the properties of continuity and decrease, nonlinear function exp(-x)[24] is selected as the basis function for similarity calculation. The similarity formula based on relative interval span is shown in (7).

$$sim(u,v)^{RIS}=\begin{cases}\sum_{i\in A}\exp\left(-|r_{u,i}-r_{v,i}|\right)\\ \sum_{i\in B}\exp\left(-|r_{u,i}-r_{v,i}|\cdot\theta\right)\end{cases} \tag{7}$$

The set A represents the same interval subgroup, and the set B represents the non-same interval subgroup.

4. Algorithm analysis

The implementation process of the similarity algorithm for sparse data is shown in Algorithm 1.

Algorithm 1: Similarity algorithm for sparse data

Enter: User-Project rating table $R(m,n)$

Output: Top-N [25] neighbor set of user u

Create a set of non-zero rated items for user u, and calculate the rating probability according to the set of non-zero rated items.

Calculate user similarity based on relative interval span according to Equation (7), and the similarity factor takes value between intervals.

Calculate user rating preference similarity according to the rating probability calculated in Equation (2) and step a.

Calculate the structural similarity of users according to Equation (6).

Substitute the similarity calculated in steps b, c and d into the similarity model equation (1) to calculate the overall similarity。

Numerical values λ were determined by multiple experiments on different data sets.

Predict the rating of user rating items.

Determine the top-n neighbor set of user U according to the similarity.

Time complexity analysis. Assuming that there are m users and n items in the data set, the time complexity of computing user preference similarity and user common rating item similarity is $O(m^2)$, therefore, the time complexity of the proposed model is $O(2\cdot m^2)$. The time complexity of KLCF and BCF is $O(n^2+m^4)$. The proposed algorithm has more advantages in terms of time complexity.

ISSN: 2664-9640

5. Comparative analysis of experimental results

5.1. The data set

Two open source datasets, ML-100K and ML-Latter-Small, were selected in this paper. The number of users, number of projects, number of ratings and data sparsity in the dataset are shown in Table 2. The data set is divided into training set and test set, with the training set accounting for 80% and the test set accounting for 20%.

Table 2 Data set properties

数据集	用户	项目	评级	稀疏度
ML-100k	943	1682	100000	93.7%
ML-Latest-Small	610	9742	100836	98.3%

5.2. Evaluation Index

During the experiment, the prediction accuracy and recommendation accuracy are evaluated. Accuracy of prediction. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used to measure the prediction accuracy of the algorithm, and the formula is as follows. Where N is the total number of predicted user evaluation items, and $p_{u,i}$ is the predicted value of user u for project i.

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} | r_{u,i} - P_{u,i} | \tag{8}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} | r_{u,i} - p_{u,i} |^2} \tag{9}$$

Accuracy is recommended. The comprehensive evaluation index (F1) is the weighted harmonic average of Precision and Recall, so F1 is used to measure the recommendation accuracy of the algorithm. The definitions of Precision, Recall and F1 are shown in Equations (10), (11) and (12), respectively. Where I_{TP} is the predicted recommended item set, and I_{TN} is the actual recommended item set for the test set.

$$Precision = \frac{| I_{TP} \cap I_{TN} |}{| I_{TN} |} \tag{10}$$

$$Recall = \frac{| I_{TP} \cap I_{TN} |}{| I_{TP} |} \tag{11}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{12}$$

5.3. The best θ and λ values

JSR is proposed to solve the problem of data sparsity. In order to improve the recommendation performance of the model, JSR is introduced θ and λ into the model as a weight parameter. Different values of θ and λ have different degrees of influence on the model. In order to obtain the weight value that plays a high role in the algorithm, the experiment is carried out on the dataset ML-latch-small with relatively high sparsity to determine the value of θ and λ.

5.3.1. Selecting the best Value of θ

To determine the value of weight parameter θ, the second part of the similarity model, namely the product of structural similarity and interval span similarity, is calculated independently. Under different neighboring conditions, the final MAE value is calculated by taking different

ISSN: 2664-9640

values for θ. The value is controlled in an interval. The experimental results are shown in Table 3.

Table 3 MAE at different θ values

	K					
λ	30	40	50	60	70	80
0.30	0.72962	0.72403	0.72102	0.71933	0.71752	0.71673
0.40	0.72921	0.72376	0.72084	0.71920	0.71733	0.71668
0.50	0.72918	0.72369	0.72060	0.71914	0.71711	0.71642
0.60	0.72920	0.72385	0.72065	0.71922	0.71713	0.71646
0.70	0.72918	0.72355	0.7209	0.71915	0.71721	0.71644

As can be seen from Table 3, under different values of θ, the MAE value decreases with the increase of the number of neighbors and gradually becomes stable. When θ value is 0.5, the MAE value is the lowest, indicating that the prediction accuracy is the highest at this time. In order to achieve the optimal effect of the overall model, the weight parameter value is selected as 0.5 (applicable to all data sets).

5.3.2. Selecting the best Value of λ

Under different neighboring conditions, by taking different values for λ, the final MAE and F1 values are calculated and compared. The λ value is controlled in the range of 0.55 and 0.75. The experimental results are shown in Table 4 and Table 5 respectively.

Table 4 MAE at different λ values

	K					
λ	30	40	50	60	70	80
0.55	0.72868	0.72298	0.72002	0.71859	0.71645	0.71566
0.60	0.72865	0.72292	0.71994	0.71850	0.71637	0.71558
0.65	0.72861	0.72285	0.71987	0.71841	0.71628	0.71550
0.70	0.72863	0.72288	0.71987	0.71843	0.71629	0.71550
0.75	0.72848	0.72272	0.71971	0.71826	0.71611	0.71528

Table 5 F1 at different λ values

	K					
λ	30	40	50	60	70	80
0.55	0.66025	0.66374	0.66572	0.66673	0.66764	0.66928
0.60	0.66028	0.66392	0.66597	0.66679	0.66758	0.66949
0.65	0.66065	0.66419	0.6662	0.66779	0.66689	0.66949
0.70	0.66050	0.66364	0.66566	0.66664	0.66682	0.66803
0.75	0.66038	0.66413	0.66575	0.66658	0.66612	0.66767

As can be seen from the table, with the increase of the value of λ, the MAE value shows a downward trend, indicating that with the increase of the weight of rating preference similarity, the prediction accuracy of the algorithm is gradually improved. And F1 value is the highest when value is 0.65, indicating that when A value is 0.65, the recommendation accuracy of the algorithm is optimal. In order to achieve the optimal recommendation and prediction accuracy of the model, the weight parameter A was selected as 0.65 according to the analysis of the results in the table.

ISSN: 2664-9640

5.4. Analysis of laboratory results of the overall model

Experiments were conducted on two datasets with different sparsity, and seven collaborative filtering algorithms, JMSD[26], RJMSD[27], NHSM[15], TMJ[18], CPCC[28], BCF[16] and KLCF[17], were selected for comparative tests. The interval of the number of neighbors is the range of 10 and 100, and the step is set to 10.

5.4.1. Verify the effectiveness of the proposed structural similarity

On the ML-latest-small dataset, the representative number of neighbors 40 and 60 are selected, and the classical structural algorithm Jaccard[12], the newer Sorensen index[29] (Srs), and the Sreepadalike-minded[30] structural algorithm are selected. It is compared with MAE, RMSE and F1 values of the proposed algorithm SJaccard to verify the effectiveness of the proposed structural similarity. The experimental results of all algorithms are shown in Table 5.

Table 6 Comparison of similarity effect of different structure types

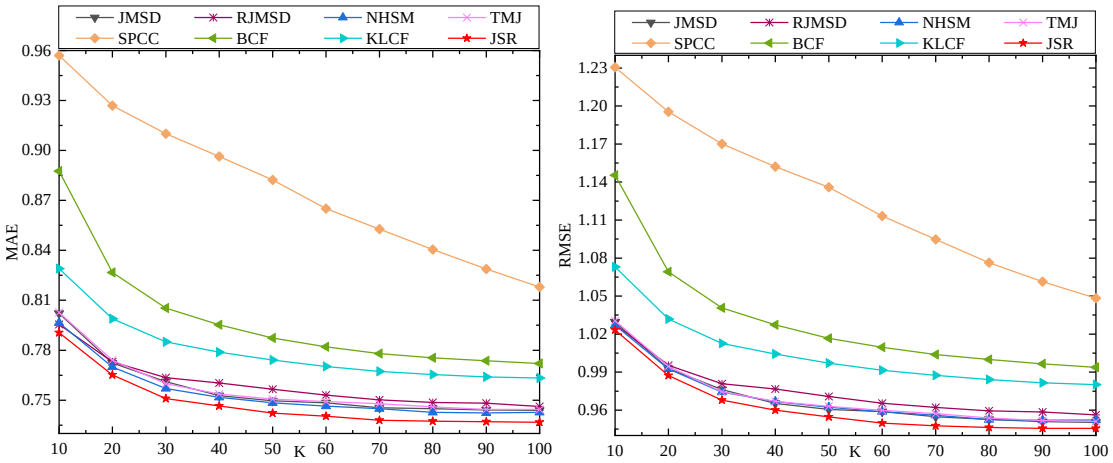
	MAE		RMSE		F1	
算法	40	60	40	60	40	60
Jaccard	0.733708	0.725992	0.954299	0.943013	0.636075	0.645253
Srs	0.733677	0.725963	0.954234	0.942939	0.636486	0.645051
Sreepadalike-minded	0.733287	0.726133	0.952478	0.942425	0.635268	0.645214
SJaccard	0.728731	0.723536	0.948606	0.940542	0.641391	0.648416

As can be seen from Table 6, SJaccad has the lowest MAE and RMSE value and the highest F1 value when the number of neighbors is 40 and 60. In terms of MAE and RMSE, SJaccard increases by about 0.4% compared with the other three structural similarity models. Based on F1, SJaccard also improved by about 0.4%. The overall prediction accuracy and recommendation accuracy are better than the comparison structural similarity algorithm. Possible causes: SJaccard fully considers the rating structure and the number of equal rating values, which can improve the similarity of users with similar ratings, increase the distinction between users, and thus improve the quality of neighbor recommendation. It shows that SJaccard has the optimal effect on the accuracy of prediction and recommendation, which verifies the effectiveness of the proposed structural similarity.

5.4.2. JSR prediction accuracy analysis

On the ML-100K dataset, the MAE and RMSE values of all algorithms varying with the number of neighbors are shown in Figure 1. In FIG. 1(a), the MAE values of all algorithms decrease with the increase of neighbor number K. Among the comparison algorithms, the MAE value of NHSM algorithm is the lowest, that is, NHSM has better score prediction ability compared with other algorithms. At the same time, the MAE value of JSR algorithm is significantly better than other algorithms under different number of neighbors, and is about 0.7% higher than NHSM, indicating that the proposed algorithm has good prediction accuracy. In FIG. 1(b), when the number of neighbors is 10, the difference between the RMSE value of JSR algorithm and NHSM and RJMSD algorithm is small, but with the increase of the number of neighbors, the difference is gradually obvious, and JSR shows the optimal score prediction performance. In the algorithm based on global rating, the MAE value and RMSE value of JSR are lower than those of JMSD, RJMSD, BCF and KLCF. In the comparison algorithm, JMSD shows the best performance, but JSR still improves 0.5% compared with JMSD algorithm. Compared with KLCF and BCF based on rating prediction, the MAE has improved by 2-3%, and the RMSE has improved by 4-6%.

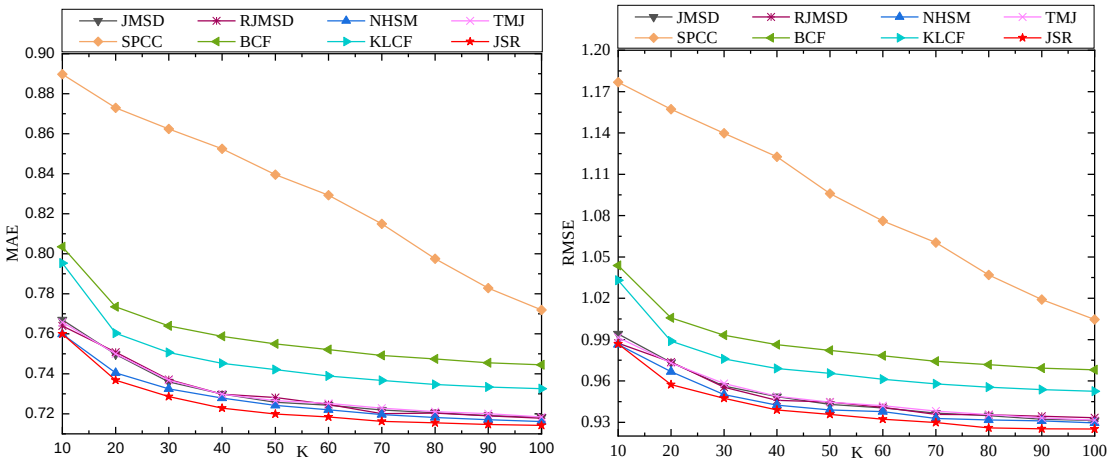
ISSN: 2664-9640



(a) MAE on ML-100K (b) MAE on ML-100K

Figure.1 MAE and RMSE on ML-100K

On the ML-latest-Small dataset, the MAE values and RMSE values of all algorithms are shown in Figure 2. Compared with the results on ML-100K dataset, NHSM and JSR algorithms have more obvious advantages in prediction accuracy, indicating that NHSM and JSR algorithms have better adaptability under sparse data, and JSR algorithm is superior to NHSM algorithm, so JSR algorithm has the best performance. Compared with KLCF algorithm, which also introduces rating probability, JSR algorithm has 1.5-3% improvement in MAE value and JSR has 3-5% improvement in RMSE value. As can be seen from the table, JSR algorithm has the best prediction accuracy. The main reasons may be as follows: Firstly, JS divergence breaks the unbounded property of KL divergence, improves the similarity accuracy when measuring rating preference similarity, and considers the impact of rating value on structural similarity to enhance user discrimination. Secondly, structural similarity improves the quality of neighbor recommendation by considering the number of equal ratings. Finally, according to the importance of rating preference similarity and numerical similarity, the overall similarity is proportionally divided to improve the reliability of the model.



(a) MAE on ML-Latest-Small (b) RMSE on ML-Latest-Small

Figure.2 MAE and RMSE on ML-Latest-Small

5.4.3. JSR recommendation accuracy analysis

All algorithms were executed on the ML-100K and ML-latests-small datasets, with F1 values shown in Figures 3 and 4. With the increase of the number of neighbors, the F1 value of JSR algorithm shows an upward trend. In FIG. 3, except when the number of neighbors is 10, the F1 value of BCF algorithm is higher than that of JSR algorithm, and all the others are lower than that of JSR algorithm. Compared with the recently proposed KLCF algorithm, the JSR algorithm improves by 1.2-1.6% and the BCF based on rating probability improves by about 1%. In other

ISSN: 2664-9640

words, JSR algorithm has a high advantage in the accuracy of recommendation. In Figure 4, JSR algorithm keeps the closest trend to the recently proposed BCF and KLCF algorithms, but JSR algorithm still has an improvement of about 0.6%. Compared with JMSD algorithm, JSR still improves by about 1.8% in prediction accuracy. It shows that the performance of JSR algorithm is stable. On sparse data sets, JSR algorithm still maintains good recommendation performance, indicating that the model has high adaptability.

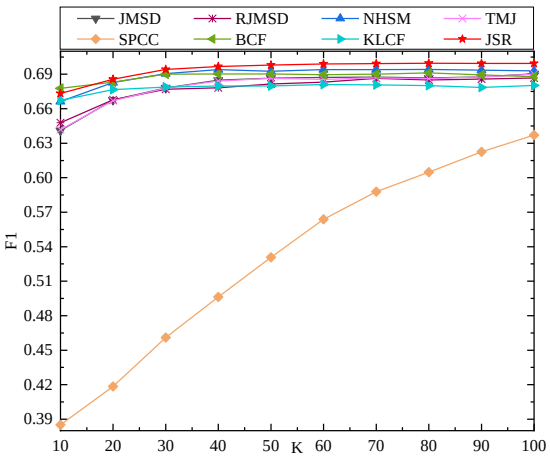


Figure.3 F1-value on ML-100K

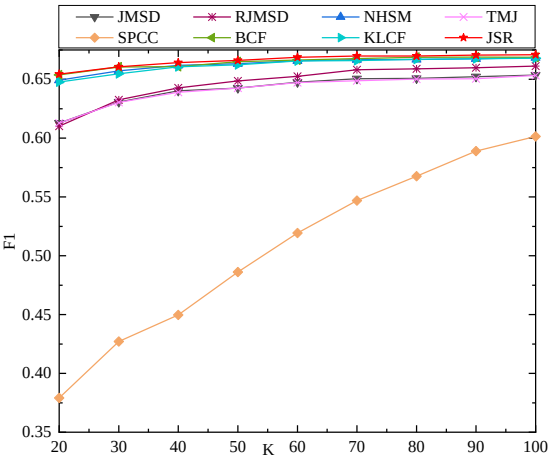


Figure.4 F1-value on ML-Latest-Small

6. Conclusion

In this paper, a similarity model for sparse data is proposed to improve the adaptability and accuracy of the algorithm under sparse data. JSR similarity model is developed from two aspects: rating structure and rating value difference, and makes effective use of global rating information. Firstly, from the perspective of user rating probability distribution and according to the rating quantity information, the improved JS divergence based rating preference similarity is proposed to alleviate the data sparsity problem and improve the similarity accuracy. In addition, SJaccard considers the role of the number of the same rating values on structural similarity, introduces the number of the same rating set information, improves the integrity of structural similarity, and emphasizes the importance of common rating from the perspective of rating structure. Finally, the common rating items of users are divided into the same interval subgroups and the non-same interval subgroups, and the differentiated similarity calculation method is designed for different subgroups, and the weighted similarity factor is defined to improve the discrimination degree of similar users, obtain the similarity degree based on interval span, and improve the performance of the recommendation algorithm. Compared with existing algorithms, JSR algorithm improves the accuracy of prediction and recommendation, and has stronger adaptability under sparse data.

References

[1] Zhou Mingyang, Xu Rongqin, Wang Ziming, et al. A generic Bayesian-based framework for enhancing top-N recommended algorithms[J]. Information Sciences, 2021,580(1):460-477.

[2] Wu Sen, Dong Yaxian, Wei Guiying, et al. Research on User Similarity Calculation for Collaborative Filtering for Sparse Data [J/OL]. Computer Science and Exploration: 1-12 [2022-05-04]. (Wu Sen, Dong Yaxian, Wei Guiying, et al. Research on user similarity calculation of collaborative filtering for sparse data[J/OL]. Journal of Frontiers of Computer Science and Technology:1-12[2022-05-04].)

[3] Huang Zhenhua, Zhang Jiawen, Tian Chunqi, et al. Review of Recommendation Algorithms Based on Ranking Learning [J]. Journal of Software, 2016, 27(03): 691-713. (Huang Zhenhua, Zhang Jiawen, Tian Chunqi, et al. Survey on learning-to-rank based recommendation algorithms[J]. Journal of Software, 2016,27(03):691-713.)

[4] Khrouf H, Troncy R. Hybrid event recommendation using linked data and user diversity [J]. 2013:185–192.
[5] Ajaegbu, Ciigozirim. An optimized item-based collaborative filtering algorithm[J]. Journal of Ambient Intelligence and Humanized Computing, 2021.
[6] Mohana H, Suriakala M. An Enhanced Prospective Jaccard Similarity Measure (PJSM) to Calculate the User Similarity Score Set for E-Commerce Recommender System[M]. 2021:129--142.
[7] Chen Ting, Zhu Qing, Zhou Mengxi, et al. Trust-based recommendation algorithm in social network environment [J]. Journal of Software, 2017, 28(03): 721-731. (Chen Ting, Zhu Qing, Zhou Mengxi, et al. Trust-based recommendation algorithm in social network[J]. Journal of Software, 2017, 28(03):721-731.)
[8] Suryakant, Mahara T, Kant S. A new similarity measure based on mean measure of divergence for collaborative filtering in sparse environment[J]. Procedia Computer Science, 2016, 89: 450-456.
[9] Resnick P. GroupLens: An open architecture for collaborative filtering of Netnews[J]. Proc Cscw, 1994.
[10] Candillier L, Meyer F, Boulle M. Comparing State-of-the-Art Collaborative Filtering Systems[C]// International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, Berlin, Heidelberg, 2007:548- -562.
[11] LIU Haifeng, HU Zheng, Mian A, et al. A new user similarity model to improve the accuracy of collaborative filtering[J]. Knowledge-Based Systems, 2014, 56:156-166.
[12] Felfernig A, Jeran M, Ninaus G, et al. Toward the next generation of recommender systems: applications and research challenges[M]//Multimedia services in intelligent environments. Springer, Heidelberg, 2013: 81-98.
[13] Jonathan, L, Herlocker. An algorithmic framework for performing collaborative filtering[J]. ACM SIGIR forum, 2017, 51(2):227-234.
[14] Ahn H J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]. Information Sciences, 2008, 178(1):37-51.
[15] Liu Haifeng, Hu Zheng, Mian A, et al. A new user similarity model to improve the accuracy of collaborative filtering[J]. Knowledge-Based Systems, 2014, 56(11):156-166.
[16] Patra B K, Launonen R, Ollikainen V, et al. A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data[J]. Knowledge-Based Systems, 2015, 82(C):163-177.
[17] Deng Jiangzhou , Wang Yong , Guo Junpeng , et al. A similarity measure based on Kullback–Leibler divergence for collaborative filtering in sparse data[J]. Journal of Information Science, 2018.
[18] Sun Shuangbo, Zhang Zzhiheng, Dong Xinling, et al. Integrating triangle and jaccard similarities for recommendation. PLoS One. 2017 Aug 17;12(8): e0183570.
[19] Zhang Hengru, Min Fan, Zhang Zhiheng, et al. Efficient collaborative filtering recommendations with multi-channel feature vectors[J]. International Journal of Machine Learning and Cybernetics, 2018.
[20] Jesus Bobadilla, Ortega F, Henando A, et al. A collaborative filtering similarity measure based on singularities[J]. Information Processing & Management, 2012, 48(2): 204-217.
[21] Wang Yong, Wang Yongdong, Deng Jiangzhou, et al. Recommendation Algorithm Integrating Jensen-Shannon Divergence [J]. Computer Science, 2019, 46(02): 210-214. (Wang Yong, Wang Yongdong, Deng Jiangzhou, et al. A recommendation algorithm based on Jensen-Shannon divergence[J]. Computer science,2019,46(02):210-214.)
[22] Khrouf, Troncy H A, Rapha. Hybrid event recommendation using linked data and user diversity[C]//Proceedings of the 7th ACM Conference on Recommender Systems, Association for Computing Machinery, 2013:185–192.
[23] Majter A P, Lamberti P W, Prato D P. Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states[J]. Physical Review A, 2005, 72(5):762-776.
[24] Gazdar A, Hidri L. A new similarity measure for collaborative filtering based recommender systems[J]. Knowledge-Based Systems, 2020, 188.
[25] Subramaniyaswamy V, Logesh. Adaptive KNN based recommender system through mining of user preferences[J]. Wireless Personal Communications, 2017, 97(2): 2229-2247.
[26] Bobadilla J, Serradilla F, Bernal J, et al. A new collaborative filtering metric that improves the behavior of recommender systems[J]. Knowledge-Based Systems, 2010, 23(6):520-528.
[27] Bag S, Kumar S K, Tiwari M K, et al. An efficient recommendation generation using relevant Jaccard similarity[J]. Information Sciences, 2019, 483:53-64.

ISSN: 2664-9640

[28] Herlocker J L , Komstan J A , Borchers A , et al. An algorithmic framework for performing collaborative filtering[J]. ACM, 1999.
[29] Pirasteh P , Hwang D, Jung J E . Weighted Similarity Schemes for High Scalability in User-Based Collaborative Filtering[J]. Mobile Networks and Applications, 2014, 20(4):497-507.
[30] Sreepada R S , Patra B K . An Incremental Approach for Collaborative Filtering in Streaming Scenarios[M]. Springer, Cham, 2018.