ORIGINAL RESEARCH PAPER

Volume-8 | Issue-8 | August - 2019 | PRINT ISSN No. 2277 - 8179

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

RESEARCH ON SIMILARITY ALGORITHM BASED ON USER BEHAVIOR CONSISTENCY

Computer Science

Zhang Haichao School of computer science and technology, Tianjin Polytechnic University, Tianjin, 300387, China

Yang Lianhe* School of computer science and technology, Tianjin Polytechnic University, Tianjin, 300387, China *Corresponding Author

ABSTRACT

Collaborative filtering algorithm is one of the most classical and successful recommendation algorithms. The similarity measurement method of traditional collaborative filtering algorithm ignores the consistency of user behaviors, which leads to inaccurate similarity calculation. Aiming at the above problems, this paper designs a similarity calculation method based on user behavior, this method rewards and penalizes the similarity calculation according to whether the user scores are consistent, and also takes into account the dispersion of user scores and the proportion of common scores among users in the similarity calculation process. The experimental results show that compared with the other four algorithms, when the nearest neighbor is 5, the accuracy of the algorithm is improved by 2.34%, 3.49%, 5.41% and 9.33%, respectively, which effectively improves the recommendation quality.

KEYWORDS

collaborative filtering, similarity, behavioral consistency, dispersion

INTRODUCTION

Personalized recommendation technology actively explores the most needed resources for users by researching the interests of different users, and better resolves the contradiction between the increasing Internet information and user need[1]. Collaborative filtering recommendation is one of the most widely used and successful technologies to date[2-4].The basic idea is to generate recommendations to target users based on the rating data of nearest neighbors with similar scores[5].

The most critical component of the collaborative filtering mechanism is to effectively discover the similarity between users, so the precision of the similarity calculation will affect the accuracy of recommendation[6]. The traditional similarity calculation method has some shortcomings. In response to the problems above, literature [7] proposed a similarity algorithm using the combination of Jaccard similarity and mean squared difference similarity. Literature [8] proposed a user similarity model that considered both local context information of user ratings and global preferences of user behavior. Literature [9] proposed a similarity measure model of normal distribution function.

This paper designs a similarity calculation method based on user behavior. The experimental results show that the proposed algorithm can effectively improve the recommendation accuracy and recall.

TRADITIONAL COLLABORATIVE FILTERING ALGOR ITHM

In a general recommendation system, we usually have *m* users, *n* items and the sparse user-item rating matrix$mnRR\times\in$. Each r_{ij} of *R* denotes the user *u*'s rating on item *i*. In the user-based collaborative filtering method, the core is to find the neighbors of the target users by calculating the similarity between users[10]. The most common method is cosine similarity. The formula is as follows:

$$sim(u,v)^{\cos}=\frac{\sum_{k\in I_{uv}} r_{uk}\times r_{vk}}{\sqrt{\sum_{k\in I_{uv}} r_{uk}^{\ 2}}\times\sqrt{\sum_{k\in I_{uv}} r_{uk}^{\ 2}}} \quad (1)$$

Where I_{uv} represents the set of items that the user *u* and the user *v* are jointly scored, and r_{uk} represents the score of the user *u* on the item *k*.
The similarity between users is obtained by the similarity calculation method, and the target user's score for the unrated item is predicted by the set *Q* of the top *N* users most similar to the target user *u*, and the formula is as follows:

$$p_{ui}=\bar{r}_u+\frac{\sum_{n\in Q} sim(u,n)\ (r_{ni}-\bar{r}_n)}{\sum_{n\in Q} sim(u,n)} \quad (2)$$

Where p_{ui} represents the predicted score of user *u* for item *I*, $sim(u,n)$ represents the similarity between user *u* and user *n*, r_{ni} represents the score of user *n* for item *i*, and*nr*represents the mean of user *n* for all items.

IMPROVED SIMILARITY CALCULATION METHOD

This paper proposes a similarity calculation method based on the consistency of user behaviors. The similarity calculation method in this paper is called PDJ, and the method contains three factors: Proximity, Dispersion, and Jaccard. These three factors measure the consistency, difference and common score ratio of the scores between users. The calculation formula is as follows:

$$sim(u,v)^{\text{PDJ}}=sim(u,v)^{\text{Proximity}}\ sim(u,v)^{\text{Dispersion}}\ sim(u,v)^{\text{Jaccard}} \quad (3)$$

Proximity

The higher the score, the more interested the user is in the item. Suppose that for one item, the average score for all users for this item is *m*. Then consider [0, *m*) as a negative score and (*m*, 5) as a positive score. Since each user's scoring habits are different, whether the user's rating of an item is positive or negative is for the average score of the user. Therefore, there are the following definitions:

$$Agr(r_{uk},r_{vk})=\begin{cases}1 & (r_{uk}-\bar{r}_u)\ (r_{vk}-\bar{r}_v)\geq 0\\ 0 & \text{otherwise}\end{cases} \quad (4)$$

When two users score the same item, if both users score higher or lower than the average of their scores, then the two users are considered to be consistent scores, that is, the value of *Agr* is 1; Otherwise, it is considered to be an inconsistent score, that is, the value of *Agr* is 0. According to the value of *Agr*, the Proximity factor is defined as follows:

$$Agr(r_{uk},r_{vk})=\begin{cases}1 & (r_{uk}-\bar{r}_u)\ (r_{vk}-\bar{r}_v)\geq 0\\ 0 & \text{otherwise}\end{cases} \quad (5)$$

The Proximity factor indicates how close the two users are to the same item. The initial value of the score is 1/2. If the scores of the two users are consistent, the two are considered to be positively close, i.e., the initial value plus the proximity of the two. If the scores of the two users are inconsistent, the two are considered to be passively close, i.e., the initial value is subtracted from the proximity of the two. Therefore, the $sim\ (u,v)^{\text{Proximity}}$ of the two users is defined as follows:

$$sim(u,v)^{\text{Proximity}}=\sum_{k\in I_{uv}}\text{Proximity}(r_{uk},r_{vk}) \quad (6)$$

DISPERSION

The closer the scores between the two users, the more similar the interests of the two users. Traditional similarity calculation method often uses variance to measure the difference of user scores. However, if the measurement scales of different users are different, it cannot directly use standard deviation for comparison, so the effects of measurement scale and dimension should be eliminated. The coefficient of variation is standardized according to the average value of the mean, and the data can be objectively compared. The coefficient of variation measures the degree of dispersion of data[12] and is defined

as follows:

$$CV_u = \frac{\sqrt{\frac{1}{I_u}\sum_{k \in I_u}(r_{uk} - \bar{r}_u)^2}}{\bar{r}_u} \quad (7)$$

Where CV_u represents the coefficient of variation of the user u.

For any two users, the larger the coefficient of variation difference, the greater the score dispersion, and the greater the difference between users. Conversely, the smaller the coefficient of variation difference, the smaller the dispersion of the two user scores, and the smaller the difference between users. Therefore $sim\ (u, v)^{\text{Dispersion}}$ is defined as follows:

$$sim(u,v)^{\text{Dispersion}} = 1 - \frac{1}{1+\exp(-|CV_u - CV_v|)} \quad (8)$$

Jaccard

When calculating user similarity, the proportion of user common scores cannot be ignored. Different from the traditional Jaccard method, the algorithm uses the improved Jaccard method. Experiments show that the improved Jaccard is better than the traditional Jaccard method. The definition is as follows:

$$sim(u,v)^{\text{Jaccard}} = \frac{|I_u \cap I_v|}{|I_u| \times |I_v|} \quad (9)$$

EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify the validity of the proposed PDJ algorithm, this section compares it with the adjusted cosine similarity, the Pearson correlation coefficient, NHSM algorithm and NDF-CF[9] algorithm.

Dataset

The dataset used in the experiment is derived from the Movielens dataset, which contains 50,000 user ratings of 100,000 movies for 1,682 movies. All scores fall within the [0, 5] range. In this experiment, the data set was randomly divided into 80% of the data as a training set and 20% of the data as a test set.

Evaluation

Precision describes how much of the final list of recommendations is the user-item scored record that has occurred, which is defined as follows:

$$\text{Precision} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|} \quad (10)$$

Where U represents a set of all users, R (u) represents a list of recommendations made to the user, and T (u) represents the user's scored list on the test set.

The recall describes how much of the user-item scored record is included in the final list of recommendations, which is defined as follows:

$$\text{Recall} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|} \quad (11)$$

Coverage describes the ability of the recommended algorithm to exploit long tail items, which is defined as follows:

$$\text{Coverage} = \frac{|\bigcup_{u \in U} R(u)|}{|I|} \quad (12)$$

Where I represents the set of test set items.

RESULT

The number of recommended experimental items is 10, in order to find the nearest neighbor number K of the target user to achieve the best experimental results, the recommended effect of Top10 is tested with 6 different K values. Each similarity calculation method repeats 5 experiments, and the average of these 5 experiments was taken as the final result. The evaluation results of Precision, Recall, and Coverage are shown in Figures 3 to 5.

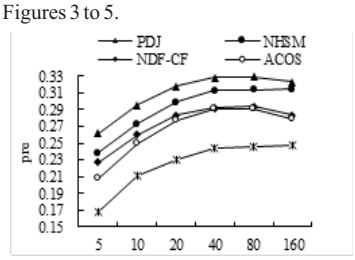
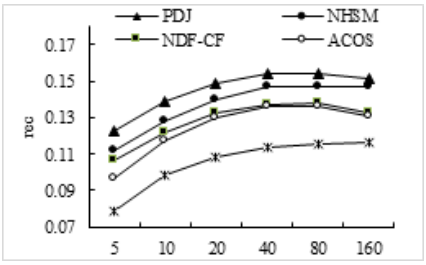
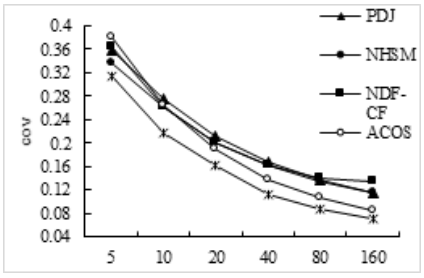


Figure 3 Precision comparison result



Figu4 Recall comparison result



Figu5 Coverage comparison result

The experimental results above show that:

1) With the increase of K value, the precision and recall of the 5 methods are increased, and when K is in the range of [0, 40], the growing speed of precision and recall is fastest. When K is in the range of (40,160), the precision and recall of all methods tend to be stable. Among them, when K=80, the precision and recall of PDJ, NDF-CF and ACOS methods reach the maximum. When K>80, the precision and recall of the three methods are reduced.
2) In different situations, the Pearson method has the lowest precision and recall. The NDF-CF method is slightly better than the ACOS method, but at K=40, the precision and recall of the two methods are very close. The NHSM method has a good performance under different K values, but the PDJ method proposed in this paper is always superior to the other 4 methods in precision and recall.
3) With the increase of K value, the coverage of the 5 methods continues to decrease. The Pearson method has the lowest coverage with different K values. When K is in the range of [0, 20], the coverage of the other 4 methods is very close. The PDJ method is better than other methods in most cases. However, when K=5, the coverage of the PDJ method is lower than that of the ACOS and NDF-CF methods. At K=160, the coverage of the NDF-CF algorithm performs best.

CONCLUSION

Aiming at the low computational accuracy of the traditional similarity calculation method, this paper proposes a similarity calculation method based on the consistency of user behaviors. The experimental results show that compared with the traditional similarity algorithm, the PDJ algorithm proposed in this paper effectively improves the precision, recall and coverage of recommendation results.

The similarity calculation in traditional collaborative filtering relies on the user-item scoring matrix, but the matrix is an extremely sparse matrix, and many users do not have a common scoring item, so that the similarity between users cannot be calculated. Therefore, it will be the next research content to solve the sparse problem of the scoring matrix.

REFERENCES

[1] Xing Chunxiao, Gao Fengrong, Zhan Sinan, et al. Collaborative Filtering Recommendation Algorithm Adapted to User Interest Changes [J]. Journal of Computer Research and Development, 2007, 44(2):296-301.
[2] Li Ling. Research on collaborative filtering algorithm based on information entropy [D]. Beijing: Beijing Jiaotong University, 2018.
[3] Greg Linden, Brent Smith, Jeremy York. Amazon.com recommendations: Item-to-Item collaborative filtering [J]. IEEE Internet Computing, 2003, 7(1):76-80.
[4] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. IEEE Trans. Knowl. Data Eng., 2005, 17(6):734-749.
[5] Deng Ailin. Zhu Yangyong. Collaborative Filtering Recommendation Algorithm Based on Project Scoring Prediction [J]. Journal of Software, 2003, 14(9):1621-1628.
[6] ZHANG Nan, LIN Xiaoyong, SHI Wei-hui. Collaborative filtering recommendation method based on improved heuristic similarity model [J]. Journal of Computer Applications, 2016, 36(8):2246-2251.
[7] BOBADILLA J, SERRADILLA F, BERNAL J. A new collaborative filtering metric that improves the behavior of recommender systems [J]. Knowledge-Based Systems, 2010,

23(6):520-528.
[8] Liu Haifeng, Hu Zheng, Mian A, et al. A new user similarity model to improve the accuracy of collaborative filtering [J]. KnowledgeBased Systems, 2014, 56(3):156-166.
[9] Qiu Guoqing, Ma Jun, Zhao Wei, et al. Collaborative filtering algorithm based on similarity of normal distribution function [J]. Application Research of Computers, 2018, 35(10):2920-2923.
[10] Zheng Yuping, Hu Minjie, Yang Honghe, et al. Research on collaborative filtering algorithm based on rough set [J/OL]. Journal of Shandong University (Science Edition), 2019, 54(1):1-10.
[11] LIU H, HU Z, MIAN A, et al. A new user similarity model to improve the accuracy of collaborative filtering [J]. Knowledge-Based Systems, 2014, 56(3):156-166.
[12] Xu Yi, Tang Yimin, Wang Wei. Collaborative Filtering Algorithm Based on Positive and Negative Correlation Nearest Neighbors [J].Engineering Science and Technology, 2018, 50(5):189-195.