International Journal of Advanced Computer Technology (IJACT)
ISSN: 2319-7900 Volume VIII, Issue II, April, 2019

SLOPE ONE COLLABORATIVE FILTERING BASED ON USER SIMILARITY

Ye Xu, Information Engineering College, Henan University of Science and Technology, Luoyang 471023, China
Haichao Zhang, Information Engineering College, Henan University of Science and Technology, Luoyang 471023, China

Abstract

To solve the problem of data sparsity in traditional collaborative filtering algorithm and improve the accuracy of recommendation algorithm, a Slope based on user similarity is proposed.

One Filling Score Matrix Collaborative Filtering Algorithms. The algorithm first uses the basic cosine similarity to calculate the similarity between users and generates the user similarity matrix, then fills the original user similarity matrix with the score predicted value of Slope One algorithm as the backfill value of the first n neighbors, and finally recommends the filled similarity matrix according to the traditional item-based collaborative filtering algorithm. Slope One collaborative filtering algorithm based on user similarity takes full account of the similarity between users when filling the matrix, which makes the score prediction more accurate. The new algorithm alleviates the problem of data sparsity to a certain extent and improves the accuracy of the algorithm. The improved algorithm, classical collaborative filtering algorithm and even-weighted Slope One algorithm are tested on Movie Lens dataset. The results show that the Slope One collaborative filtering algorithm based on user similarity effectively alleviates the problem of data sparsity and has better recommendation effect.

Introduction

Today's society is an era of information explosion, which is filled with various kinds of information. The problem of "information overload" makes it difficult for users to find the information they need in a short time, resulting in a waste of time. On the other hand, with the development of society, people's pursuit is varied and different, and the problem of how to meet the unique needs of each person remains to be solved. In this context, personalized recommendation technology emerged and widely used in major video websites, e-commerce, mobile media, such as amazon, taobao, Jingdong mall, Dangdang and so on.

The personalized recommendation algorithms used in each system are also different. The commonly used recommendation algorithms are content-based recommendation algorithm, user-based recommendation algorithm, Combination-Based recommendation algorithm, Association rule-based recommendation algorithm and collaborative filtering-based recommendation algorithm [2]. Among them, the recommendation algorithm based on collaborative filtering is one of the most widely used and most effective recommendation algorithms. However, with the increasing number of users and projects, data sparseness, inaccurate similarity calculation has become a key factor affecting the performance of recommendation system. In this regard, scholars at home and abroad have done a lot of research. Literature [3] Proposed a recommendation algorithm combining project clustering and Slope One scheme. Using the project clustering algorithm, the project is clustered into several clusters, and Slope One algorithm is applied to several clusters to predict the unknown project score of the target user. Reference [4] proposes a recommendation algorithm which combines project clustering and Slope One scheme. Item clustering algorithm is used to aggregate items into several clusters, and Slope One algorithm is applied to each cluster to predict the target user's score on unknown items. Literature [5] The Slope One algorithm is used to calculate the score prediction value to fill in the data, and based on the filled data, the similarity is modified and the nearest neighbor selection set is optimized. Finally, the recommendation list of the target user is given. Li et al. [6] In order to solve the problem of sparse data and low accuracy in recommendation system, a collaborative filtering algorithm is proposed, which combines user rating trust and user preference trust to improve score similarity. Literature [7] proposes a collaborative filtering algorithm based on matrix clustering. The user rating data are clustered by matrix clustering algorithm. Then the collaborative filtering algorithm is applied to the clustered sub-matrices, which improves the recommendation accuracy of the algorithm.

The recommendation results of these algorithms show that the quality of recommendation system has been improved, but there are still sparse data and recommendation quality problems in collaborative filtering algorithm. In the data filling algorithm, the filling form is single and the filling item score is inaccurate, which leads to low confidence and does not consider the similarity between users, thus affecting the accuracy of the recommendation results. In order to solve this problem, this paper first calculates the similarity between items by improved cosine similarity calculation method, and generates user similarity matrix from the first N items with the highest similarity. Then, Slope One filling algorithm is used to fill the similarity matrix, and the predicted results are backfilled into the original user-item score matrix. User-based collaborative filtering algorithm is used to recommend the results, which can alleviate the problem of data sparsity to a certain extent and improve the accuracy of the predicted values.

A Brief Introduction to Collaborative Filtering Recommendation Algorithms

In 1992, Goldberg, Nicols, Oki and Terry first proposed the concept of collaborative filtering. Collaborative filtering recommendation algorithm is the earliest and well-known recommendation algorithm. The algorithm discovers user's preferences by mining user's historical behavior data, divides users into groups based on different preferences and recommends similar items. It is a typical method of using collective wisdom. Its basic idea is to recommend items of interest to users.

Collaborative filtering algorithms are divided into model-based collaborative filtering recommendation and memory-based collaborative filtering recommendation [9-11]. Memory-based system filtering algorithm is divided into user-based system filtering and Project-based Collaborative Filtering algorithm, which is mainly divided into the following steps: user rating-matrix is used to calculate user similarity to find users with similar interests and generate similarity matrix, and then the Unrated items of users to be recommended are predicted according to the neighbor rating, and finally the results are obtained according to the rating. The results were recommended.

Collaborative filtering recommendation algorithm is the earliest and well-known recommendation algorithm. The main function is prediction and recommendation. The algorithm discovers users ‘preferences by mining users' historical behavior data, divides users into groups based on different preferences and recommends products with similar tastes. Collaborative filtering recommendation algorithms are divided into two categories: user-based collaborative filtering algorithm and item-based collaborative filtering algorithm. Simply put: people gather in groups, and things in groups.

A. Similarity calculation

Several basic methods of similarity calculation are based on vectors, that is, calculating the distance between two vectors, the closer the distance is, the greater the similarity is. In the recommended scenario, in the two-dimensional user-preference matrix, we can use a user's preferences for all users as a vector to calculate the similarity between users, or use all users ‘preferences for an item as a vector to calculate the similarity between items. Here are some commonly used similarity calculation methods.

Euclidean Distance

It was originally used to calculate the distance between two points in Euclidean space. Assuming that X and y are two points in n-dimensional space, the Euclidean distance between them is:

$$d(x,y)=\sqrt{\left(\sum\left(x_i-y_i\right)^2\right)} \qquad (1)$$

As you can see, when n = 2 is, the Euclidean distance is the distance between two points in the plane.

When Euclidean distance denotes similarity, the following formulas are generally used for conversion: the smaller the distance, the greater the similarity.

$$sim(x,y)=\frac{1}{1+d(x,y)} \qquad (2)$$

Pearson Correlation Coefficient:

Pearson correlation coefficient is generally used to calculate the degree of tightness between two fixed-distance variables, whose values are between [-1,+1].

$$p(x,y)=\frac{\sum x_i y_i - n\overline{xy}}{(n-1)s_x s_y}=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^{\;2}-(\sum x_i)^2}\sqrt{n\sum y_i^{\;2}-(\sum y_i)^2}} \qquad (3)$$

It is the standard deviation of X and Y samples.

Cosine Similarity:

Cosine similarity is widely used to calculate document data similarity:

$$T(x,y)=\frac{x\bullet y}{\|x\|^2\times\|y\|^2}=\frac{\sum x_i y_i}{\sqrt{\sum x_i^{\;2}}\sqrt{\sum y_i^{\;2}}} \qquad (4)$$

Tanimoto Coefficient:

Also known as Jaccard coefficient, it is an extension of Cosine similarity and is also used to calculate the similarity of document data:

$$T(x,y)=\frac{x\square y}{\|x\|^2+\|y\|^2-x\square y}=\frac{\sum x_i y_i}{\sqrt{\sum x_i^{\;2}}+\sqrt{\sum y_i^{\;2}}-\sum x_i y_i} \qquad (5)$$

a. Score prediction

The k-nearest neighbor (12) method is usually used for scoring prediction, i.e. K-nearest neighbor sets which are most similar to the target users are selected as the nearest neighbor sets for calculation. Assuming that the set u represents the nearest neighbor set of the target user u, the predicted score of user u for item I is:

$$p(u,i)=\bar{R}_u+\frac{\sum\limits_{u_k\in U} sim(u,u_k)\times(R_{uk,i}-\bar{R}_{u,k})}{\sum\limits_{u_k\in u} sim(u,u_k)}$$

(6)

After completing the scoring prediction, n users who have the highest scoring and are not in the target user's evaluated item set are selected as top-N recommendation set to implement the recommendation.

B. Model Design of Slope One System Filtering Algorithms Based on User Similarity

a. Model description

The process of the algorithm is mainly divided into two parts: 1) Score matrix filling module: According to the original user-item score matrix, the similarity between users is calculated preliminarily, and the user-item score matrix of similar users is formed. The first n users with high similarity are selected to fill the similarity matrix with slope one algorithm compared with other empty value filling algorithms. The prediction of the algorithm is based on slope one algorithm. Values are more reliable. 2) Recommendation module: Predict the items that users have not scored, scoring the items that users have not evaluated by item-based scoring prediction on similarity matrix, average the results of two scoring, and recommend the first n items in the set as top n recommendation set.

b. Algorithm flow

Based on the above related concepts and computational models, the algorithm flow described in this paper is as follows:

1. Input the original sparse "user-item" score matrix.
2. For each one: (calculating the similarity between each user)
3. Generating k-Nearest Neighbor Sets of Users
4. Slop one Filling Similarity Matrix
5. Prediction of items not scored by users based on similarity matrix
6. Considering the similarity between items, the item-based collaborative filtering algorithm is used for filling matrix, and the items that are not scored are predicted.
7. Find the mean of the two calculation methods, and take the first n items as top n recommendation set to implement recommendation.

C. Simulation experiment and result analysis

a. experimental data

The simulation uses the MovieLens data set [13] provided by the GroupLens project team to evaluate the algorithm. The data set contains 943 user information, 1682 project information, and 10,000 user rating information. Each user has at least 20 rating records. The higher the rating range is, the greater the user's preference for movies. In the experiment, 80% of the score data are used as user training model of training set, and the remaining 20% are used as test set. Five groups of data were randomly selected from the data set and the data were not intersected. The cross-validation method was used to evaluate the algorithm.

b. Metric standard MAE

Common evaluation and recommendation system accuracy indicators include user satisfaction, prediction accuracy, diversity, surprise, novelty, real-time, robustness, trust, business goals. At present, prediction accuracy, Top N recommendation and coverage are the main methods to evaluate offline experiments. Mean Absolute Error (MAE) is used to measure the recommendation quality of the algorithm. MAE is a commonly used method for users to measure the accuracy of statistics and compare the quality, which can accurately reflect the quality of recommendations. It is used to measure the error between the predicted user rating and the actual user rating. The smaller the MAE value, the higher the recommendation accuracy. Assuming that the user interest set predicted by the system is (p1, p2... pn), and its actual interest set is (q1, q2... Qn), the calculation method is as follows:

$$MAE = \frac{\sum_{i=1}^{n} |p_i - q_i|}{n} \tag{7}$$

c. Analysis of experimental results

In order to analyze the influence of the main parameters in the algorithm on the recommendation effect, the number of neighbors is mainly analyzed.

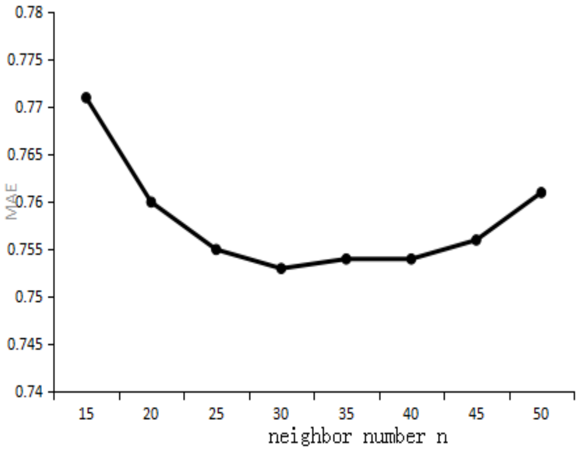

Figure 2 MEA values vary with neighbor n

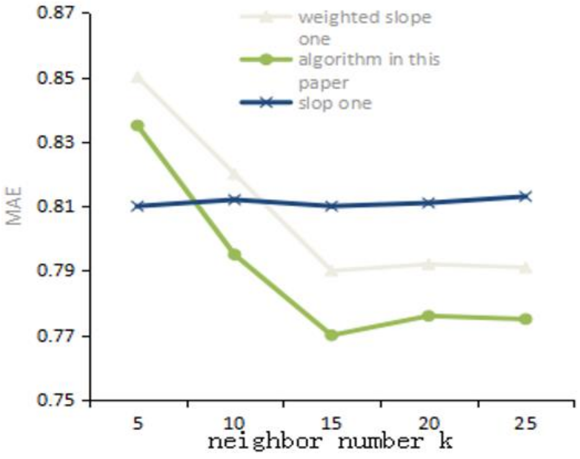

Figure 2 MAE values vary with neighbor K

In order to verify the effectiveness of the proposed algorithm, the MAE values of the proposed algorithm and the existing algorithms are compared under the same parameter environment. When using the improved slope one algorithm to fill in the evaluation of non-scored items, it is necessary to select user ratings which have high similarity interest preferences with the corresponding users of the current items to be scored. That is to say, it is necessary to find the nearest user of the current user in the user preference matrix. In this experiment, the range of the number of nearest neighbors n is 15-50, and then observe the change of n to recommend the algorithm accurately. The influence of degree and the value of n with good recommendation effect are selected. The experimental results are shown in the figure. Experiments show that when the number of nearest neighbors is around 30, the MAE value of the prediction score of this algorithm reaches the lowest, that is, the recommendation effect of the algorithm is the best, and the quality of the recommendation is the best. Therefore, the number of nearest neighbors is 30 in this paper.

In order to evaluate the recommendation accuracy of the proposed algorithm, this paper compares it with the other two algorithms, including the traditional slope one algorithm, the improved weighted slope one algorithm and the proposed algorithm. The experimental results show that the three algorithms can be improved to some extent with the increase of the number of neighbors K. Compared with other two algorithms, the proposed algorithm can reduce the MAE value and improve the accuracy of algorithm recommendation. It can be seen that the collaborative filtering algorithm and slope one algorithm based on fusion and user have better accuracy and can effectively alleviate the problem of data sparsity.

Conclusions

In this paper, a user-based collaborative filtering algorithm and a slope one-based matrix filling algorithm are proposed to solve the problem of recommendation quality degradation caused by data sparsity. The slope one algorithm is used to predict the non-scored items of the user similarity matrix, and then the original user-item scoring matrix is filled back. On this basis, the user-based collaborative filtering algorithm is used to implement the recommendation. The experimental results show that the proposed algorithm can effectively improve the recommendation quality.

References

[1]. Gouws R H, Tarp S Information overload and data overload in lexicography [J]. International Journal of Lexicography, 2017: ecw030.

[2]. Herlocker L, Konstan A, Borcher S A, et al. An algorithmic framework for performing collaborative filtering [C]//Proc of, International ACM SIG IR Conference on Research and Development in Information Retrieval.1999:230-237.

[3]. You Haipeng, Li Hui, Wang Yunmin, et al. An improved collaborative filtering recommendation algorithm combing item clustering and slope one scheme [C]. Lecture Notes in Engineering &amp; Computer Science, Vol 2215.2015:313-316.

[4]. You Haipeng, Li Hui, Wang Yunmin, et al. An improved collaborative filtering recommendation algorithm combining item clustering and slope one scheme [C] Lecture Notes in Engineering &amp; Computer Science, vol2215 2015:313-316.

[5]. Xiangdong, Qiu Zixian. Research on collaborative filtering algorithm based on slope one algorithm to improve score matrix filling [J/OL]. Computer application research, 2019 (05): 1-5 [2018-09-21].

[6]. LI Liang. Dong Yuxin, Zhao Chunhui, et al. Collaborative filtering recommendation algorithm combined with user trust [J] Journal of Chinese Computer Systems, 2017, 38 (5) 951-955

[7]. Gao Fenglong, Xing Chunxiao, Du Xiaoyong, etc. Collaborative filtering algorithm based on matrix clustering [J]. Journal of Huazhong University of Science and Technology. Natural Science Edition, 2005, 33 (S1): 257-260 (Gao Fengrong, Xing Chunxiao, Du Xiaoyong, Wang Shan A collaborative filtering algorithm based on matrix clustering [J] Journal of Huazhong University of Science and Technology (Natural Science Edition), 2005, 33 (S1): 257-260)

[8]. Goldberg D, Nichols D, Oki B M, et al. Use collaborative filtering to weavr an information tapestry [J]. Communications of the ACM. December, 1992.35 (12): 61-70

[9]. Dehghani Z, Reza S, Salwah S, et, al. A systematic review of scholar context aware recommender systems [J] Expert Syst Appl 2015 (42): 1743

[10]. Adomavicius G, Tuzhilin A Toward the next generation of recommender system: a survey of the state-of-art and possible extensions [J].IEEE Trans on Knowledge &amp; Data Engineering, 2005, 17 (6): 734-749

[11]. MinhA, Salakhutdinow R. Probability matrix factorization [C]//Advances in Neural Information Processing Systems 2007:1257

[12]. Rosin, Ouyang Yuanxin, Xiong Zhang, etc. optimize collaborative filtering algorithm based on K-nearest neighbor through similarity support [J]. Journal of Computer Science, 2010, 33 (8): 1437-1445.

[13]. MovieLens_100k [DB/OL]. https://grouplens.org /datasets /movielens/.

Biographies

YE XU born in 1993, is a master student of Information Engineering Institute ,Henan University of Science and Technology Information Engineering, with interest in Personalized recommendation algorithm. Ms Xu may be reached at xuye_0723@163.com.

Haichao Zhang born in 1963. He is a professor at Henan University of Science and Technology. He is mainly engaged in the research of graphic image processing and computer control.