Research on Collaborative Filtering Recommendation Algorithm Based on Matrix Decomposition Method

LiJuan

Longnan Teachers' College, Chengxian 742500,China

Keywords: Matrix decomposition, Collaborative filtering, Data mining, Least squares, Personalization, Regularization.

Abstract. In order to realize the personalized recommendation of internet mass data, according to the characteristics of internet mining data set and combined with mathematical algorithms, this paper proposes a new forecasting and computing model of adding the regularization constraint and least square method based on the traditional matrix decomposition model (SVD), improving the speed and accuracy of the proposed algorithm. Matrix decomposition before and after improvement carries out experiments and results analysis with filtering recommendation algorithm, the experimental results show that the speed and accuracy of two prediction score calculation methods have some promotion after adding the regularization constraint and the least squares. After joining the regular constraints, the RMSE values obtained by MATLAB will monotonic decrease, avoiding the over fitting phenomenon and improving the calculation quality.

Introduction

In today's Internet Era, the amount of information is growing at a constant geometry. Internet users have begun to not worry about the lack of information, but rather to worry about how to get the effective information [1,2]. Search engine is an important tool to obtain Internet data, and it also changes the situation of information explosion to a certain extent, improving the user to obtain the ability of target data in mass data [3]. However, the search engine is in the construction of key words, there are still many problems in the information extension and novelty. Personalized recommendation system can use data mining calculation, exploring the data relevance according to the relevant mathematical algorithms and providing interested content for the users, which has important significance for the mass of Internet data processing.

Information Explosion and Data Mining Recommendation Algorithm

Internet has brought a lot of information for people, but also caused the problem of information explosion, the amount of information is too large, and people do not know what information useful, which is useless, so we are unable to know what information we want [4-6]. Twitter information released ninety million in a day, the average number of users on the Facebook reached 130, while youtube uploaded video per minute also reached an average of 34. Internet has experienced a number of major experiences.

Portal era. In the portal era, the main activity of the internet is the portal site traffic, and people can query their own needs in the various categories of the portal. The portal information is provided primarily by the portal's professionals, it is similar to the traditional media and newspapers, users receives information passively because they cannot create information and provide feedback.

Search engine era. With the increasing amount of information on the internet, the demand of people has also changed, the user begins to receive information from passive to active search information, so the search engine came into being, it is an effective way to deal with the information explosion, and the most representative is Google.

Web2.0 era. In the Web2.0 era, everyone has become the internet information provider. Internet changes to read and write by read only, the supply of information has soared, including blog, SNS, Wiki and other typical internet applications. Web2.0 exacerbated the impact of information explosion,

and the question of search answer reached a million, so that we cannot respond for the diversity of search results.

In the vast amount of search data, we need to use personalized recommendation system. According to the corresponding algorithm, we carry out data mining and machine learning, the use of internet users' ratings and feedback information fully tap the data features, and users can really get the data in the mass data.

Collaborative Filtering Recommendation Algorithm based on Improved Matrix Factorization

In order to improve the matrix decomposition and filtering algorithm, the traditional algorithm is analyzed in the recommendation system [7]. Suppose that the matrix of *m* users and *n* rating objects is *R*, the recommended object feature matrix and user feature matrix are respectively *V* and *U*. After collaborative filtering algorithm based on SVD is simplified, the input is the user's score matrix *R* and the characteristic number *d*, output is the approximation matrix *X* of the matrix *R*, in which the specific steps of the algorithm can be summarized as follows [8-10]:

(1) Data initialization, the user can obtain R_{norm} after rating matrix *R* carries out standardization;
(2) Determining the matrix dimension *k*, *S* is simplified as *k* dimension matrix, so as to get S_k;
(3) After using the SVD algorithm decomposes R_{norm}, we can get *U*, *S* and *V*;
(4) In accordance with the steps (3), U_k and V_k are simplified;
(5) Calculating the square root S_k, it will be recorded as $S_k^{1/2}$;
(6) According to $S_k^{1/2}$, the calculation of the relevant matrix $S_k^{1/2}V$ and U$S_k^{1/2}$;
(7) For the *j* prediction score, the use *i* can be written $P(i, j) = \overline{R_i} + U_k S_k^{1/2}(i) S_k^{1/2} V_k(j)$.

By using the above algorithm, the user can get the prediction score for any project, in which $\overline{R_i}$ is the average value in all evaluation item users. In order to improve the algorithm, we can find a low rank matrix *X* to carry on the maximum extent approximation on *R*. The minimize Frobenius loss function is

$$L(x) = \sum_{ij} (R_{ij} - X_{ij})^2 \tag{1}$$

Among them, *L(x)* shows the objective function, $(R_{ij}\text{-}X_{ij})^2$ shows the square error term of low rank approximation, and the improved algorithm can be solved quickly on *L(x)*. In the recommendation system, in order to realize the algorithm, the formula (1) is rewritten

$$L(U,V) = \sum_{ij} (R_{ij} - U_i V_j^T)^2 \tag{2}$$

In order to prevent the fitting problem in the calculation process, the formula (2) is a regularization constraint, and the formula (2) can be rewritten as

$$L(U,V) = \sum_{ij} (R_{ij} - U_i V_j^T)^2 + \lambda(\|U_i\|^2 F + \|V_j\|^2 F) \tag{3}$$

V is fixed, we can solve U_{i} formula on U_{i} derivation, it is

$$U_i = R_i V_{ui} (V_{ui}^T V_{ui} + \lambda n_{ui} I)^{-1}, i \in [1 \quad m] \tag{4}$$

Among them, R_i shows the film composition vector after users *i* are rated, V_{ui} shows the user evaluation of film feature vector matrix, n_{ui} indicates the number of film comment. Similarly, V_i will be fixed and derivate, it can get

$$V_j = R_j U_{mj} (U_{mj}^T U_{mj} + \lambda n_{mj} I)^{-1}, j \in [1 \quad m] \tag{5}$$

Among them, R_i represents the film composition vector after user *j* rating, U_{mj} represents the user *i* film review feature vector matrix, n_{mj} represents the number of film reviews. In formula (4) and formula (5), *I* shows d*d unit matrix, the regularization cross can use the least square algorithm to continue to optimize. First of all, we use the Gauss random number initialization matrix of mean

value 0 and deviation 0.01, and then the use of formula (4) updates *U*, the use of formula (5) updates *V*, until the calculation value of RMSE is convergence, or the number of iterations is enough, the calculation can stop.

The Experimental Results and Analysis of Improved Recommendation Algorithm

This algorithm uses MATLAB software to process the MovieLens data set, and the data set is created and maintained by the GroupLens research group of America Minnesota University [11]. The data packet contains 100000 scoring record of 900 users for the 1600 films.

Experiments are divided into three groups that are traditional SVD algorithm, joining regularization constraint algorithm, joining regularization constraint algorithm and least squares algorithm experiment [12,13]. Each algorithm uses the configuration same computer, the computer is in the same local area network. In the experiment, the number of iterations is 80, and the number of features is 1-9. The results obtained by calculation are shown in Figure 1.

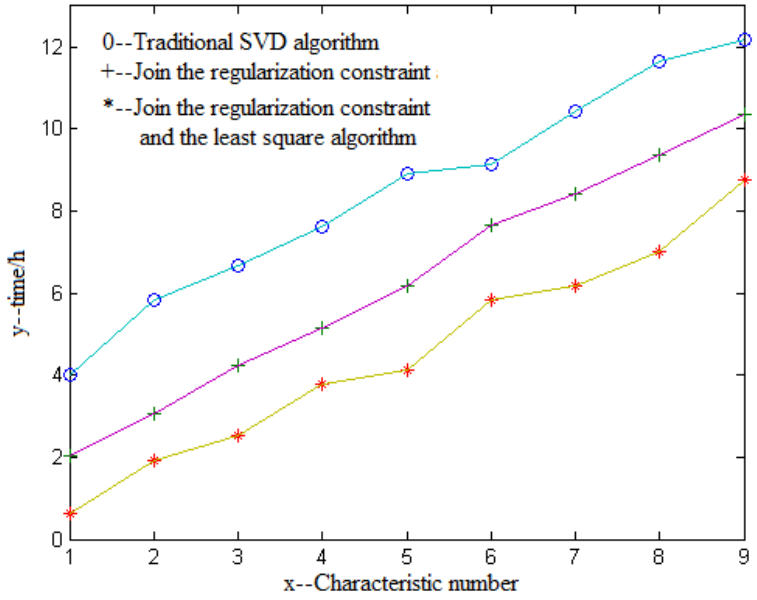

Fig. 1 The running time comparison of different algorithms

Figure 1 shows the comparison results of the three experimental group running time, in which the horizontal axis is the number of features, the vertical coordinates is the running time. The graph can be seen that after adding a regular constraint and the least squares algorithm, running time is significantly shortened, and with the increase of the number of features, this effect is more obvious.

Table 1 shows the accuracy of different algorithms, it can be seen that the improved algorithm accuracy is obviously higher than the traditional algorithm accuracy, and simultaneously adding the regular and least square algorithm are more accurate than join the algorithm accuracy of regular constraint alone.

Table 1. Comparison of the traditional SVD algorithm and the improved algorithm calculation accuracy

Experiment number	Traditional SVD algorithm calculation accuracy	Join regularization constraint calculation accuracy	Join the regular and the least squares method calculation accuracy
1	0.913	0.952	0.992
2	0.912	0.956	0.995
3	0.925	0.962	0.987
4	0.918	0.953	0.996
5	0.9163	0.966	0.988

In order to further study the effect of the algorithm, the results of the experiments use RMSE evaluation criteria, RMSE goes through the calculation forecast deviation between user score and actual score to measure the accuracy of the forecast, which is most commonly measurement method of recommended level. The smaller the value of RMSE, the higher the quality of the recommendation, hypothesis that the prediction score vector of N project is $\{p_1,p_2,\cdots,p_N\}$, and the actual user's score vector is$\{r_1,r_2,\cdots,r_N\}$, the RMSE calculation of the algorithm is

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(p_i-r_i)^2}{N}} \tag{6}$$

Using MATLAB software calculates the traditional algorithm and the RMSE value of improved SVD algorithm, and we can get the calculation of the final results as shown in Figure 2.

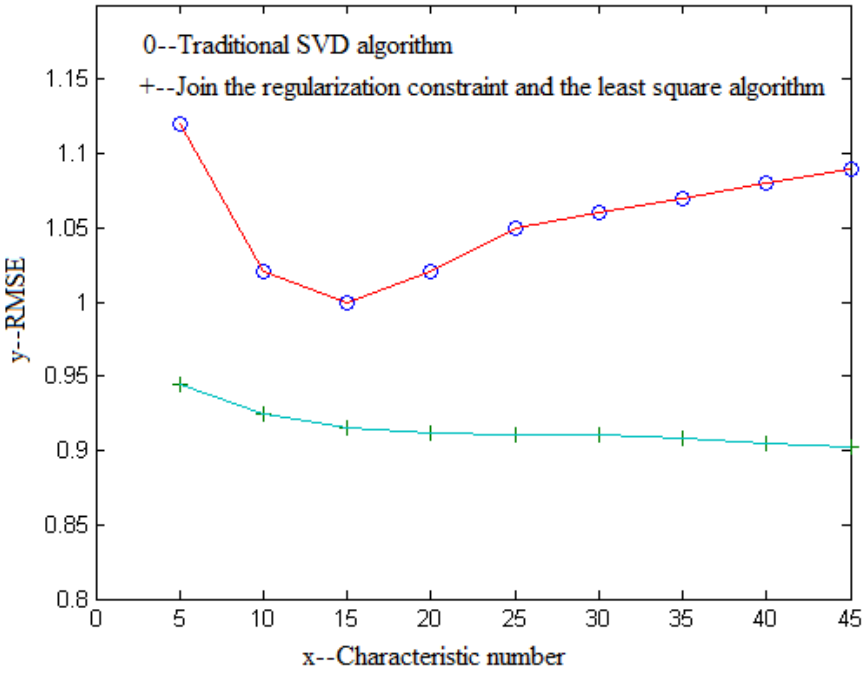

Fig. 2 The performance comparison of improved algorithm

Figure 2 shows the algorithm performance comparison before and after the improvement, the algorithm can be seen that the improved algorithm amplitude is better than the original algorithm amplitude under each characteristic. With the increase of the number of features, the RMSE value of the improved algorithm monotonic decreases, the feature matrix *U* and *V* are constrained after the reason of the decrease is mainly due to the addition of the regularization constraint, so as to avoid the occurrence of over fitting phenomenon. However, the traditional algorithm is not added constraint, and RMSE will be increased with the increase of the number of features, so the fitting phenomenon has been affected the quality of the calculation.

Summary

According to the characteristics of internet search engine data mining, this paper joins the regularization constraint and the least square algorithm based on the traditional matrix decomposition model (SVD) and designs a new internet data personalized recommendation system. In order to verify the reliability of the system, before and after the improved algorithm is carried out experiments and results analysis, the experiment uses MATLAB software on MovieLens data processing. Under the calculation of different characteristic numbers, the time and precision of different algorithms are calculated by MATLAB iterative calculation. The experimental results show that before and after the improved algorithm can effectively improve the speed and accuracy of prediction score calculation, which can get a better RMSE value.

References

[1] X. Xu, X.F. Wang. Analysis of the cheat and attack behavior of collaborative filtering algorithm based on SVD. Computer engineering and applications, 2014, 45(20): 95-97.

[2] X. Luo, Y.X. Ouyang, Z. Xiong. The collaborative filtering algorithm based on K nearest through the similarity support optimization. Journal of computer science, 2013, 33(8): 22-26.

[3] Q. Wang, L.R. Zheng. The collaborative filtering recommendation algorithm based on common score and similar weight. Computer science, 2013, 37(2): 38-45.

[4] G.L. Sun, H.L. Qi. The spam filtering based on online scheduling logic regression. Journal of Tsinghua University, 2013, 53(5): 734-740.

[5] B.T. Liu. Research on data mining algorithm based on rough set. China West technology, 2013, 10(14): 11-12.

[6] J.C. Hu, G.P. Wu. Improved genetic BP neural network data mining algorithm and its application. Micro machine and application, 2013(2): 30-34.

[7] B. Chu, C. Wu, X.B. Yang. Data mining algorithm based on RBF neural network and rough set. Computer technology and development, 2013, 23(7): 87-91.

[8] Q.P. Yang, Y. Ding, Y.M. Qian. Research on data mining platform and its key technology based on cloud computing. ZTE technology, 2013, 19(1): 53-60.

[9] B. Huang, S.R. Xu, W. Pu. Design and implementation of data mining platform based on MapReduce. Computer engineering and design, 2013, 34(2): 495-501.

[10] Q.S. Yu. The design and implementation of logistic regression model algorithm based on cloud platform. Technology Bulletin, 2013, 29(6): 137-139.

[11] Z.M. Gu, J.X. Zhang, C. Zheng. Overview of cloud computing progress research. Computer applications, 2014, 27(2): 429-433.

[12] Y. He, W.Q. Wang, F. Xue. Research on massive data mining based on cloud computing. Computer technology and development, 2013, 23(2): 69-72.