Research on User-based Normalization Collaborative Filtering Recommendation Algorithm

Jie Dong[1, a ,*], Jin Li[1,b], Gui Li[1,c], Liming Du[1,d]

[1]Information and Control Engineering College, Shenyang Jianzhu university, Shenyang, 110168, China

[a]email:dong_jie_2002@163.com, [b]email:951802281@qq.com [c]email:ligui21c@sina.com , [d]email:wangfy678@163.com, [*]corresponding author

Keywords: Normalization, Collaborative Filtering, Sparsity, Recommendation Algorithm

Abstract. Under the circumstance of the big data, because of the low efficiency and low performance of analysis and calculation in stand-alone mode, the traditional recommendation algorithm is limited greatly, recommended time and recommended precision is difficult to guarantee. This thesis makes a improvement on the user-based collaborative filtering algorithm. The user similarity calculation is based on the normalized method, which makes user rating more reasonable and reduces the data sparse. Meanwhile the algorithm can be run on the massive clusters, which improve the operation efficiency of the system and the scalability of the system significantly. Finally, designing the scheme of recommendation system experiments and the experimental results show that the accuracy of the improved algorithm is superior to the traditional collaborative filtering algorithm and strengthen the efficiency at the same time.

Introduction

Big data greatly increases data sets of the recommendation system. The traditional recommendation system in monolithic mainframe environment is limited by its hardware, which cause many problems such as long calculation time and poor scalability and so on, it become the bottleneck of its development[1-3]. Research focusing on the recommendation system has changed from stand-alone environment to server cluster [4-6].

In 2011, Chin Feng Lai put forward to achieve the personalized recommendation by Map Reduce distributed parallel computing framework and run k-mean algorithm depending on the Hadoop environment [7].In 2012,. Lee Gai and others realize the parallel collaborative filtering algorithm based on ALS on Hadoop platform, significantly improving the operation efficiency of collaborative filtering algorithm based on ALS[8]. At present, collaborative filtering recommendation methods are mainly based on matrix decomposition [9][10]. However, the current method in scalability performance is poor, and unable to adapt to the massive data processing.

In this paper, the traditional user-based collaborative filtering recommendation algorithm is analyzed deeply, aiming that the recommendation system accuracy is not high under the circumstances of the data sparseness problem,. We improved the algorithm and run on the platform of Hadoop Map Reduce computing framework implementation what has made the effect more effectively.

User-based Collaborative Filtering Recommendation Algorithm

The basic principle of user-based collaborative filtering recommendation algorithm is calculating the similar users groups to the current user by user preference information. Namely when a user A need a personalized recommendation, other users with similar interests can be found, then those items would be recommended to A that his similar users like but user A has never used. The key is user similarity calculation.

In order to calculate user similarity, Cosine similarity calculation is often used to calculate the degree of similarity between two users. In addition to the modified cosine similarity and Pearson correlation coefficient of similarity are often used.

Regarding the user rating of the item as a vector of n dimensional space, cosine similarity set the projects with no point to 0 and cosine similarity measure method is the cosine of the Angle between the vector obtained by calculation. Assuming $\vec{i}$ and $\vec{j}$ stand for n dimension vector of user i and user j respectively, the similarity between the users i and j is defined as follows:

$$sim(i, j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\| i \| * \| j \|} \tag{1}$$

In practice, the user rating scale has differences, but the cosine similarity did not take into account user ratings subjective preference. modified cosine similarity solves the above problem by subtracting the user average score for the item in order to reduce the subjectivity of the differences. Assuming users i and j collection scoring items in $I_{i,j}$ together, I_i and I_j are scores set of users i and j respectively, $R_{i,c}$ or $R_{j,c}$ respectively represent score of user i or user j on the item c, $\overline{R}_i$ and $\overline{R}_j$ stand for users i and j average scores respectively, the user similarity between user i and user j is as follows:

$$sim(i, j) = \frac{\sum_{c \in I_{i,j}} (R_{i,c} - \overline{R}_i)(R_{j,c} - \overline{R}_j)}{\sqrt{\sum_{c \in I_i} (R_{i,c} - \overline{R}_i)^2} \sqrt{\sum_{c \in I_j} (R_{j,c} - \overline{R}_j)^2}} \tag{2}$$

Pearson correlation coefficient [11] reflects the degree of linear correlation between two variables, scopes (-1, 1)., formula is as follows:

$$sim(i, j) = \frac{\sum_{c \in I_{i,j}} (R_{i,c} - \overline{R}_i)(R_{j,c} - \overline{R}_j)}{\sqrt{\sum_{c \in I_{i,j}} (R_{i,c} - \overline{R}_i)^2} \sqrt{\sum_{c \in I_{i,j}} (R_{j,c} - \overline{R}_j)^2}} \tag{3}$$

The Pearson correlation coefficient similarity only consider those users that give scores on the same item at the same time, without considering common grading numbers' influence on the similarity between the users, and if there is only one common between two user ratings of items, the similarity can not be calculated.

By means of similarity calculation results, the most nearest neighbors of target user could be made of the top N users which similarity value are very similar to the target user similarity degree. After that, we can obtain recommend results for the target user.

Supposing N_u is a nearest neighbors set to target user , $P_{u,c}$ is a score set of user u on the item c, calculating method as shown followed:

$$P_{u,c} = \overline{R}_u + \frac{\sum_{v \in N_u} sim(u, v) * (R_{v,c} - \overline{R}_v)}{\sum_{v \in N_u} | sim(u, v) |} \tag{4}$$

Through the above methods, scores could be predicted on those items which users never use, recommend the object could be feedback to the user in accordance with the final prediction score from high to low choice of N results.

Improved UC - BC Algorithm

the improvement of computing user similarity. The formula (1)-(4) show that the user similarity computing depends on the score on items given by users. In fact, different users have different grading standard. Although modified cosine similarity considered this problem, but it still slightly rough. In this paper, we introduced the subjective normalization algorithm. By using the initial users-item evaluation matrix A (m, n) to calculate the normalized scores of user and to obtain a

new user-item evaluation matrix A’ (m, n) to eliminate the error that caused by user inconsistent standards. The specific implementation process is shown below.

Step1, according to the score matrix, calculate each user u average score r_u, calculation method is as follows:

$$\overline{r_u} = \frac{\sum_{i=1}^{n} r_{u,i}}{n} \tag{5}$$

In the formula, $r_{u,i}$ indicates the score given by the user u to the item i.

Step2, set user ratings interval by calculating averages of the minimum and maximum of all user scores as both ends of the interval , and calculation method is as follows:

$$b_{i1} = \min(r_{ui,j1}, r_{ui,j2}....., r_{ui,jn}) \tag{6}$$

$$b_{i2} = \max(r_{ui,j1}, r_{ui,j2}....., r_{ui,jn}) \tag{7}$$

$$k = \frac{\sum_{i=1}^{m} b_{i1} + \sum_{i=1}^{m} b_{i2}}{2m} \tag{8}$$

In the above formula, b_{i1} and b_{i2} are the minimum and maximum values by calculating each user the boundary value , k is average for all users interval .

Step3, recalculate score $r'_{u,i}$ that the user u give the item I in order to get a new user -item evaluation matrix $\mathrm{A}'(\mathrm{m},\mathrm{n})$

$$r'_{u,i} = \frac{k * r_{u,i}}{\overline{r_u}} \tag{9}$$

$$\overline{\overline{r_u}} = \overline{k}$$

In formula, $\overline{\overline{r_u}}$ is known as the normalized coefficient.

By the formula (9), via the derivation transformation， average score of user u has become interval average k, namely: $\overline{r'_{u1}} = \overline{r'_{u2}} = ... = \overline{r'_{um}} = k$

Using this normalization processing on the user's score, it can unify the criteria of the evaluation, which can reflect the location of the item in user’s mind. And it can effectively reduce impact which the different user rating scale on the similarity of computing projects.

When we calculate the user similarity by the traditional collaborative filtering algorithm, it can calculate on the item set that the user has scored together. If two users give the more same type of items, however the grade is less for the same item. They are similar to each other in theory, but according to the traditional similarity calculation formula, it will leads to the dissimilarity results between two users. Therefore, on the basis of the subjective normalization method, prediction score calculation method is further introduced. When we calculate the similarity between two users, we can increase user’s common grading evaluation scores to reduce data sparse and improve the quality of recommendation.

The concrete implementation process of improved filtering recommendation algorithm is shown below.

Step1, according to user-item evaluation matrix A (m, n), we are able to calculate union sets of item evaluation among each users ,calculation method as follows:

$$I_{i \cup j} = I_i \cup I_j \tag{10}$$

Among the formula: I_i and I_j are sets that users i and j have evaluated scores respectively, $I_{i \cup j}$ is a union set between the user i and j.

Step2 using subjective normalization algorithm to construct normalized score matrix, calculation

method is as follows:

$$A(m,n) \to A'(m,n)$$

Step3 according item set $I_{i\cup j}$ of users i and j and after users-item evaluation matrix normalized calculation $A'(m,n)$ to calculate similarity, the calculation method of the improved formula(3) as follows.

$$sim(u_i,u_{j)} = \frac{\sum_{c\in I}(\mathrm{r}'_{u_i,\mathrm{c}}-k)(\mathrm{r}'_{u_j,\mathrm{c}}-k)}{\sqrt{\sum_{c\in I}(\mathrm{r}'_{u_i,\mathrm{c}}-k)^2\sum_{c\in I}(\mathrm{r}'_{u_j,\mathrm{c}}-k)^2}} \tag{11}$$

Among the formula: k is the normalization factor, I represents union set that the user i and user j have evaluated items commonly by the formula (3)

Step4 assumes the user u_a is the target user, c is for forecasting project, improve user's similarity calculation formula (4) to calculate the forecast scores of user u_a on item c, calculation method is as follows:

$$P_{u_a,c} = k + \frac{\sum_{u\in U} sim(u_a,u)(r_{u,c}-k)}{\sum_{u\in U}|sim(u_a,u)|} \tag{12}$$

Test Results

In order to prove the effectiveness of the improved user-based collaborative filtering recommendation algorithm, an experiment is needed. Experimental data is adopted from the Movie Len data set provided by the online video site. The data set contains two compressed files ,that is ml–1m.zip and ml-10m.zip respectively.

In order to measure the effect of the recommendation algorithm, the average absolute deviation MAE [12] is usually used for measuring, by calculating the deviation between the score in the forecast and the user's actual score to measure the accuracy of the prediction, the smaller value of MAE means the higher the quality of recommendation.

p_i is one of the actual score of the recommended data set , q_i is the corresponding predicted score , N is the number of test data set , and MAE is defined as follows:

$$MAE = \frac{\sum_{i=1}^{N}|p_i - R_i|}{N} \tag{13}$$

The improved method and traditional method are adopted for collaborative filtering recommendation experiment respectively ,the results are shown in fig. 1.

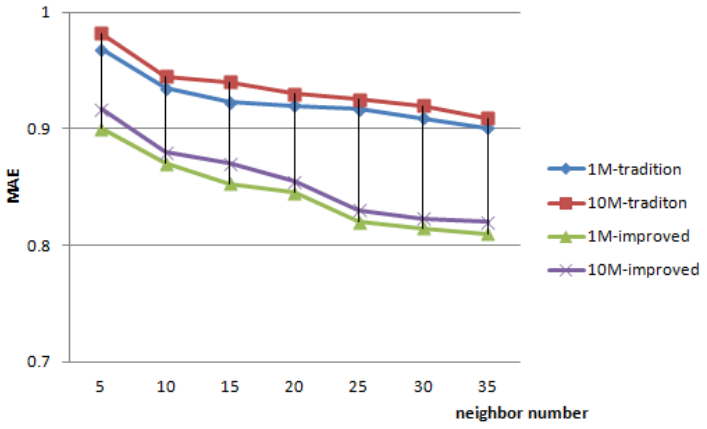

Fig1. the efficiency of different recommendation algorithm

Experimental results show that the MAE of improved collaborative filtering recommendation based on users is lower than the traditional algorithm. The sparse of 10M data sets is more than that of 1M. Using the improved algorithm, it still get better recommendation results. The improved algorithm can effectively increase accuracy of the recommendation and the efficiency of recommendation, reduce the impact of the sparse data.

Conclusion

Aiming to the problem of that efficiency is not high and the expansibility is relatively poor based on big data environment, using Map Reduce distributed programming framework on Hadoop platform, it can improve the operation efficiency and scalability of the algorithm. At the same time, by introducing the subjective normalization method to eliminate the difference of subjective criteria and the impact of computing similarity. Further, the use of improvement item similarity prediction method reduce the influence of sparse data.

Experimental result shows that the improved collaborative filtering recommendation algorithm on MapReduce framework can effectively improve the accuracy and efficiency of recommendation and increase user satisfaction.

Acknowledgement

In this paper, the research was sponsored by the Nature Science Foundation of Liaoning Province (Project No. 2014020068)

References

[1]Yuanzhuo Wang, Xiaolong Jin. Network data: the status quo and prospect [J]. Journal of Computer, 2013, 36(6):1126-1129.

[2]Jianguo Liu, Tao Zhou, Binghong Wang. The research progress of personalized recommendation system [J]. Progress in natural science, 2009, 19(1):4-10.

[3]Shahabi C, Chen Yishin. An adaptive recommendation system without explicit acquisition of user relevance feedback[J], Distributed and Parallel Databases, 2003, 14(2):173-192.

[4] Zhiyou Zhang. The overview of computer cluster technology[J]. Laboratory research and exploration, 2006, 25(5):607-609.

[5] Li Guojie, Cheng Xueqi. Research status and scientific thinking of big data [J]. Bulletin of Chinese Academy of Sciences, 2012, 27(6):647-657.

[6]Ming-Sheng Shang, Linyuan Lü, Wei Zeng, Yi-Cheng Zhang, Tao Zhou. Relevance is more significant than correlation: Information filtering on sparse data[J]. EPL(Europhysics Letters), 2009, 88(6):68008-68010.

[7] Chin-Feng Lai a, Jui-Hung Changa, Chia-Cheng Hub, Yueh-Min Huanga, Han-Chieh Chaoc. CPRS: A cloud-based program recommendation system for digital TV platforms[J]. Future Generation Computer Systems, 2011,27(6):823-835.

[8] Gai Li, Rong Pan, zhangfeng Li, Lei Li. Collaborative filtering algorithm based on large data sets of parallel study [J]. Computer engineering and design, 2012, 33(6):2437-2441.

[9] E Haihong, Song Meina, Li Chuan, et al. A collaborative filtering recommendation algorithm with time context for learning interest mining[J]. Journal of Beijing University of Posts and Telecommunications, 2014, 37(6):49-53.

[10] Ren Xiuchun, He Yaji. Research on network customer classification based on decision tree [J]. Electronic design engineering, 2014, 22 (5): 20-22.

[11] Hofmann T. Collaborative filtering via Gaussian probabilistic latent semantic analysis[C]. In: Proc. of the 26th Int'l ACM SIGIR Conf. New York: ACM Press, 2003. 259−266.

[12] Melville P, Mooney RJ, Nagarajan R. Content-Boosted collaborative filtering for improved recommendations. In: Proc. of the 18th National Conf. on Artificial Intelligence. Menlo Park: American Association for Artificial Intelligence, 2002. 187−192.