Research and improvement of the collaborative filtering recommendation algorithm

AiLing-Duan[1a] JianFeng-Wu[2b]

[1]College of Information Science and Engineering Henan University of Technology,Zhengzhou China

[2]College Of engineering training center Henan University of Technology,Zhengzhou China

[a]duanailing2006@126.com, [b]bysj320@!26

key words: Data sparsity, cloud computing, collaborative filtering recommendation algorithm

Abstract: In order to solve the user rating data about data sparseness and traditional similarity calculation method because of its disadvantages of strict match object attributes, the paper combined with the project classification and cloud computing platform is put forward an improved collaborative filtering recommendation algorithm. First of all, according to the classification of the project get class matrix; Cloud model is then used to calculate the similarity between classes within the project and get the neighbor with the highest similarity scores of the project, to score in the class project to predict filling; Using cloud model to calculate the similarity between user within the class to get users to neighbors, finally gives the final prediction score and recommendations. The experimental results show that the algorithm not only effectively solve the data sparse and the insufficiency of traditional similarity method, but also improved the user's interests and the accuracy of the nearest neighbor search; At the same time, the algorithm only need to calculate where new users or categories, greatly enhance the scalability of the system.

Introduction

The task of data mining determines the direction of data mining, and guides the algorithm to find what kind of data model. The steps of data mining will vary according to different mining tasks and different application areas, in general, can according to the characteristics of the work, the data mining tasks are divided into two major categories: descriptive data mining and prediction of data mining. The task of describing mining is to find out the general characteristics of the data, to describe the data that already exist in the database, including the generalization, the summary data, the search for the relationship between the data, the type and so on. Predictive mining is based on the current data to make inferences, in order to add more or new data to predict.

Current predictions recommendation algorithm is filtering recommendation algorithm based on content, collaborative filtering recommendation algorithm, based on demographic recommendation algorithm, based on knowledge of the recommendation algorithm and hybrid recommendation algorithm, which together filtering algorithm is one of the most used one of successful recommendation algorithm. Hadoop is a distributed memory and parallel computing based cloud computing platform, using low cost PC equipment large clusters, to build the next generation of high performance distributed mass data computing platform, are completely open source code architecture belongs not only to completely free mode, is convenient for secondary development and customization platform. With its dual advantages of high capacity and low cost, has become the driving force behind the development of large data industry, is currently the most widely used cloud computing platform. However due to the large business site users and goods in the number of large and increasing. Also, the user is given goods of scoring little, usually below 1% (5 ~ 9), resulting in data sparse user item rating matrix, seriously affect the quality of the recommendation system, coupled with the inherent cold start and scalability problems of the traditional algorithm. So the method of solving the problem of sparsity, cold start and extension is also appeared. Parallel improved collaborative filtering algorithm is proposed in this paper to solve the problem of traditional algorithms difficult to extend, for the collaborative filtering recommendation under the huge amounts of data provides a solution, has a certain reference significance.

Collaborative filtering algorithm based on the user

Collaborative filtering algorithm based on user is the basic idea is through user behavior records, and to obtain the information which can reflect the user's interest. Then the data are analyzed to produce a score matrix that reflects the relationship between the user and the product. According to the score matrix, calculate the similarity between users, so as to find similar users. For each user, you can use their similar user's interest to predict the interest of the goods, thereby achieving product recommendation. The steps of the algorithm mainly has three stages.

The user information collection.user reviews into $m \times n$ matrix; build user - item rating matrix r (m, n) represents a user rating of projects, where m is the number of users, n is the number of projects. formula (3-1) is a simple user ratings matrix in which each row represents a user rating of projects, and each column represents scores of different users of the same project, which the user item did not score the default value of 0.

The between users the calculation of the similarity.the existing basic methods are based on the vector (Vector), the existing several basic methods are based on the Vector (Vector), actually also is to calculate the distance of the two vectors, the closer distance, the greater the similarity. In the recommended scene, in the two-dimensional matrix of user preferences - items, we can be a user preference for all items as a vector to calculate the similarity between the user, or will all user preference for some items as a vector to calculate the similarity between the items. Commonly used similarity method with cosine similarity (cosine similarity), modified cosine similarity, Pearson similarity, and other methods. Each of the formula has certain difference. In practice, mainly to see how much the amount of data. Large amount of data, then calculated the similarity of the closer.

Pearson Correlation Coefficient the Correlation analysis of the Correlation Coefficient r, respectively for X and Y based on space vector is calculated after their overall standardized cosine of the Angle. Formula(3-2) is as follows:

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \\ r_{41} & r_{42} & r_{43} & r_{44} \end{bmatrix} \quad \text{（3-1）}$$

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

Sx, sy is the sample standard deviation of x and y.

Generate nearest neighbor and forecast. Will the rest of the user according to the similarity from large to small to sort and find the k most similar users as recommended by the people nearest neighbor; to k nearest neighbor similarity and nearest neighbor score on the project value prediction value of the user evaluation of the project.

To determine the user u nearest neighbor, have top - k mode and setting threshold value in two ways. The first way , Looking for k nearest neighbor of the user u, just put the Wu from big to small order, before the k for her neighbors. Generally adopts the heap sort, complexity of O (m log k). Set a threshold value is set a minimum similarity, the similarity in the Wu is greater than the lowest similarity of users as neighbors of u. Threshold and the selection of k value, largely determines the accuracy of the recommendation, on the premise of considering the precision, also need to consider the execution efficiency. Set specific values, depends on the data set. Through many experiments, to determine the specific value.

User u prediction score on project i $P_{u\ i}$, the simplest method is to use the k nearest neighbors u $_{neighbour}$ to the project i mean score formula (3-3), but this way predicted results is not good enough, did not consider similarities with your neighbors. Now commonly used is to consider the user similarity forecast method (formula 3-4), compared with considering the difference of user ratings.

$$p_{u,i} = \frac{\sum_{u_k \in u_{neighber}} r_{u_k,}i}{k} \qquad (3\text{-}3) \qquad\qquad p_{u,i} = \frac{\sum_{v \in u} w_{u} r_{v}}{\sum_{v \in u} \left| w_{u,v} \right|} \quad (3\text{-}4)$$

According to characteristics of algorithm, three problems in traditional algorithm. Data sparseness:

sparsity problems. On large e-commerce sites, Each user comments by the number of not more than 1% of the total number of projects. In this case, the score matrix is extremely sparse, there will be two have similar users by common grade project little and similarity to zero, this situation is called neighbor transmission loss. Collaborative filtering algorithm mainly rely on users to recommend resources score, score a lack of will affect the similarity between the user of the project, causing the wrong recommended.

The cold start problem. (cold - start) is also a classic problem of collaborative filtering algorithm, divided into new users, the new project. New user problem refers to a system for users joined system without any project of score can't calculation for its neighbor, also cannot be recommended for its projects. New project problem refers to just add new project due to the lack of scoring can't recommend to the user.

The scalability problems. Similarity calculation and nearest neighbor search complexity, the highest and most time-consuming. As can be seen from the Figure 1, the server side under the recommended all computing tasks. The most common solution is to improve the performance of the server, or distributed clustering is used to increase the processing capacity.

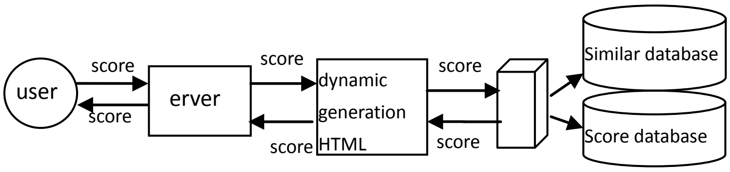

Fig 1 server working principle

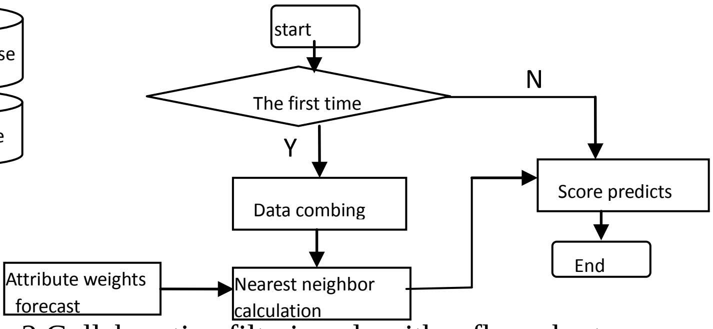

Fig 2 Collaborative filtering algorithm flow chart

Parallel improved collaborative filtering algorithm for Hadoop platform

Only for data sparseness improvement is not enough, the collaborative algorithm also faced the extension. To solve the problem below. Using the Hadoop has a strong processing capacity, and can be parallelized code only needs to implement Map and Reduce class. Begin from collaborative filtering algorithm, analysis algorithm process, the traditional collaborative filtering algorithm is decomposed into data, nearest neighbor, score, attribute weights of forecasting the four parallel Job tasks, Job order between the two, but each Job is independent of the task. Over a period of time, can be thought of as the user's interest is the same, users increased by a small amount of scoring record in the short term will not affect the calculation of similar adjacent. Therefore data combing, similarity computation, the weights of attributes to predict the three jobs do not need to perform every time.

As can be seen from the figre 2, parallel collaborative filtering can be divided into two parts of the online and offline. Combing offline part includes data, similarity calculation, attribute weights of prediction of three parts. Combing the main data processing to comb of huge amounts of data, implement classified by user id; Nearest neighbor is computing similarity between users, nearest neighbor; Property rights value forecast was primarily based on user ratings records, calculating the user on a particular attribute weights, based on the attribute value prediction is not score score value of the project. Online part is score predicts, predict users never score score value of the project.

data to comb. Under the huge amounts of data, don't put all the data in the memory. No correlation between the data, the data can be assigned to different execution node, in turn, improve the execution efficiency. According to hadoop programming process, the comb into the Map data, Combine, Reduce three stages.

Map stage to accept the file block, according to the line read each record, to deal with each record, user id, item id, the score values. After processing, the user id as the key, project id and score value as the value, the output <key, value>.

Combine stage is in order to alleviate the Map inter-node communication burden, the large amounts of data generated in the same node number data by a hyphen is connected with a common key, to Reduce function after partial results are obtained. Is the same user id as the key, project id and score value as the value, the output < key, value >. Reduce stage will be treated as multiple Map data integration processing, each all of the user's score values. Is the same user id as the key, the project id and score value as the value.

similarity algorithm is described.On Hadoop platform, the similarity calculation allocated more child nodes, use the heap sort of nearest neighbor. Nearest neighbor can use original collaborative filtering algorithm calculates the cosine value of between users or Pearson, value, or similarity and attribute weights based on user prediction algorithm to calculate the nearest neighbor, and need to calculate in advance each user preference for attribute weights. Calculate preference weight can also through the Hadoop platform parallel implementation.

Top K nearest neighbor to determine the minimum heap by building a capacity of K, by adjusting the heap of maximum K nearest neighbor. Particular way, the first design three classes to implement the neighboring calculation. Comb design SimilarityDriver class is a data entry, first of all, the Job SimilarityDriver class is initialized, set up the Map, Reduce the concrete implementation of the operation. Results can be set by adding the input file directory function, the output directory. Second, and Reduce design SimilarityMapper class make Map node is communication through the network to realize the interaction, in order not to cause the network transmission delay is too large, as far as possible, Reduce the transmission of data. So in the design of function, the computing process to a Map function will be as much as possible. In SimilarityMapper class, not only to calculate a particular user similarity with other users, also calculate topK nearest neighbor. At last, through design SimilarityReducer class Map generated k nearest neighbor, and write the final result to HDFS, calls for score predicts.

Score prediction algorithm.A grade prediction mainly through similar adjacent values to forecast of users never rated items, score predicts the nearest neighbors in the nearest Job has been calculated. Nearest neighbor specified in the score predicts Job initialization phase calculation Job output paths, in the Map phase computing user u have not rated items j to score, in the Reduce phase output prediction score values. On request, can give more child nodes will score predicts request load, so it can achieve good response to the user's request. Specific design two ForecastMapper and ForecastReducer. Through design ForecastMapper class implements the map process, according to the reading of records, analysis the user id and project id. After get the user id, read the user from a second Job output path u set corresponding to the nearest neighbor, and also need to read from the first Job score value of the nearest neighbor for project j. Secondly by designing ForecastReducer class to predict the score value output to the HDFS file. System according to the output rating value for project recommend users.

The weights of attributes. Combing the original data classified according to the user id, assign each user's scoring record child nodes perform the Map operations. In the operation of the Map to weight training of users, in Reduce operating output each user attributes weights. As before design two BWAPredictMapper and BWAPredictReducer, through BWAPredictMapper class implements the map process. Record user ratings classified by Id, and the average user assigned to compute nodes for user weight training. Finally through BWAPredictReducer class just weights of the user information integration, and then output to the HDFS.

Conclusion

Using the Hadoop cluster system consisting of six computer interconnection experiment. In order to verify the algorithm's extensibility, adopt the method of a closed TaskTracker, one by one to reduce computing nodes is to change the processing power of clusters. This experiment data size: 1.88 M, 23.4 M, 254 M. The results show that based on Hadoop collaborative filtering algorithm has good expansibility, when processing node in the cluster number increase, the shorter the time, when the data set, the greater the increase significantly.

Aiming at the shortcomings of traditional collaborative filtering algorithm, put forward specific measures for the improvement of traditional collaborative filtering algorithm. Articles with big data simulation experiments, the experimental results show that the improved algorithm on the recommended recommend efficiency and precision has obvious advantages. With the development of personalized recommendation, at the request of real-time and complexity of recommendation algorithm will be the focus of the recommendation algorithm research in the future.

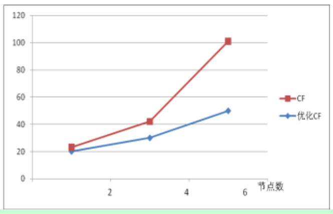

Fig 3 Collaborative filtering algorithm comparison chart

Reference

[1]Yan wei. The data mining algorithm based on cloud platform research and implementation [D].Cheng du: University of electronic science and technology，2013-03-01.

[2] Wang zhenyu,li Guo. An Analysis of the Search Engine User Behavior sBased on Hadoop.COMPUTER ENGINEERING & SCIENCE.2011.33(4).

[3huyu, Feng jun. Distributed Search Engine Using Hadoop.COMPUTER SYSTEM & Application.2010.19(7).

[4]Hang bin,Xu shuren. Design and implementation of MapReduce-based data mining platform.[J]. COMPUTER ENGINEERING AND DESIGN.2013.34(2)495-501.

[5]Taojun,Zhang ning, Collaborative Filtering Algorithm Based on Interest-Class.[J]. The computer system application.2011.20(5)55-59.

[6] duo xuesong,zhangjing etc. A Mass Data Management System based on the Hadoop. COMPUTER INFORMATION.2010.26(5-1)