A Brief Analysis of Collaborative Filtering Algorithm Based on Scoring Statistical Prediction

Liu Hongying

Xingzhi College of Xi'an University of Finance and Economics, Xi'an, Shaanxi, 710038

Keywords: scoring statistics, prediction, collaborative filtering algorithm

Abstract: With the development of the Internet and WEB technology over past years, a scoring statistical prediction model can be proposed given the item similarity and the recommended precision, so as to define users and item information. Based on this, a linear regression forecasting model is built and a corresponding algorithm is devised. In view of this, the paper discusses and analyses the collaborative filtering algorithm based on scoring statistical prediction.

With the Internet develops rapidly, it is difficult for users to process a variety of information. While recommender system can filter some information and properly select information needed. And collaborative filtering algorithm is a recommender system built on this basis, which can select some precise information for easy processing according to the preferences of users. Among the collaborative filtering algorithm, there is a scoring prediction algorithm, which grades and predicts in the light of differences among items or users rather than filtering them. In order to combine the advantages of these two algorithms and improve the precision of the recommender algorithm, the paper proposes a collaborative filtering algorithm based on the scoring statistical prediction.

1. Collaborative filtering algorithm

In the era of Internet, personalized recommendation application has become an important way for people to acquire information. While collaborative filtering algorithm is a rather successful and widely-used personalized recommendation technology based on the hypothesis: If some users get similar grades for some items, then, they will get similar grades on other items. By utilizing statistical technology to search for the nearest neighbor as well as the item results, the grades of users for projects will be predicted, and a recommendation list will be produced, which is the basic thought of the collaborative filtering algorithm. In some sense, the input data are the rating matrix of users in the recommender system (see Table 1). According to implementation method and item type, scoring methods are chosen for different users.

Table 1. Rating matrix

	$Item_1$	···	$Item_j$	···	$Item_n$
$User_1$	$R_{1.1}$	···	$R_{1.j}$	···	$R_{1.n}$
...	...	...	...	...	...
$User_i$	$R_{i.1}$	···	$R_{i.j}$	···	$R_{i.n}$
...	...	...	...	...	...
$User_m$	$R_{m.1}$	···	$R_{m.j}$	···	$R_{m.n}$

2. Analysis of collaborative filtering recommendation algorithm based on scoring prediction

The Internet has provided a strong support for big data mining, however, complicated mining algorithms require various Internets. There are redundant disks, expenses on reading and writing, and multiple applications for resources in many jobs, which hinders the performance of algorithms based on scoring statistical prediction. The collaborative filtering algorithm processed by big data,

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

relying on its advantages in iterative computing and memory computing, is able to dispatch complex computing tasks, and avoid disk accessing and resource application of intermediate results, which is suitable for data mining algorithm.

2.1 Introduction of algorithm

The development of Internet leads to information explosion. in face of mass information, how to select and filter information, select information that users focus on and feel interested in has become an issue to be solved urgently. While recommendation system, by utilizing connection between users and information, helps users acquire useful information, and on the other hand, displays interesting information for users, realizing win-win between information providers and users.

Collaborative filtering recommendation algorithm is the most classic and the most common recommendation algorithm. It finds similar users for specified user by analyzing user's interests, and forms preference prediction for a piece of information of the specified user after integrating remarks of similar users for that information. And this algorithm can be divided into three types.

User-based CF: Based on the collaborative filtering of users, the similarities among users are evaluated according to the item scores of different users in order to provide recommendation.

Item-based CF: Based on the collaborative filtering of items, the similarities among items are assessed in light of the item scores of different users in order to provide recommendation.

Model-based CF: The model-based collaborative filtering is to predict and recommend by the model created by historical materials.

2.2 Problem description

Input data format: Uid, ItemId, Rating (the scores of Uid for ItemId).

Output data: The top N ItemId with the highest similarities among each ItemId.

Due to the length limitation, the paper only selects collaborative filtering algorithm based on Item.

2.3 Algorithmic logic

It is assumed that in the item-based collaborative filtering algorithm, two similar items are quite possible to obtain favorable comments from the same user. Therefore, the algorithm first calculates the preference of users for objects. Then, based on this, similarities among items are calculated, and top N items will be found with the highest similarities. And here are detailed descriptions of the algorithm.

Calculate users' preferences: Since there may be a large gap between scores of different users for an item, each user's score should be first dualized. For example, if the score of an item given by a user is greater than the average score, then, it will be marked as 1 representing a favorable comment, otherwise, it will be marked as 0 on behalf of a negative feedback.

Calculate similarities among items: Jaccard coefficient will be used as a method to calculate the similarity between two items. Narrowly, Iaccard similarity is suitable for calculating the similarities between two sets, whose method is to use the intersection of two sets to be divided by their join, which can follow the three steps.

(1) Collect the number of favorable comments as well as the number of user making favorable comments for each item.

(2) Collect the number of keys of favorable comments as well as the number of users making the same favorable comments for two related items.

(3) Compute the similarities among items, and calculate the similarities between two related items.

Find top N items with the highest similarities. In the process, the similarities of items shall be integrated after normalization to find the top N items that are most similar to each other. And it can be divided into three steps.

(1) Normalization of similarities among items

(2) Integration of scores of similarities among items

(3) The first N items will be got with the highest similarities of each item.

2.4 The computing of algorithm

In the scoring statistical prediction model, a collaborative filtering recommendation algorithm is proposed, whose function is to calculate the personalized prediction scoring, scoring prediction of universality of items, and scoring prediction. With the gradual increase of number of items and users, extreme data will emerge. For instance, in the e-commerce system, compared with total number of items, the number of items users commenting is less than 1%, and the items commented by two users are more less. In view of this situation, traditional collaborative filtering algorithms should be changed in two ways: First, automatically classify and process according to the information. Second, acquire information like video, figure, graph, based on scoring statistical prediction and input information. In term of users' scoring characteristics and purchasing habits, the information is classified so as to ensure the similarity of item scores made by users as much as possible. In each class, the scoring information for items of users will generate a virtual user automatically, which represents the typical scores of users for the item. Then, it will serve as a new searching space to inquire the nearest neighbors and produce recommendation results.

2.5 Problems to be solved

There are several problems to be solved in the collaborative algorithm. First, in the early stage of collaborative filtering algorithm, since the common comments of users are rather less, the precision of the nearest neighbors will be reduced in some degree, failing to achieve the ideal recommendation. Hence, in the database of users' scores, due to less data, the precision of recommendation is low. Second, some new items have not obtained scores, which will impact the precision of the recommendation. And the following is the analysis of the data with less comments. Take the movie rating as an example, as for two users who like comedies, they will be calculated as no common comments by the traditional similarity calculation methods. However, if some users have no similarity but common comments with them, their similarities must be more than 0. Thus, in the initial stage of the websites with less data about user's comments, it is impracticable to calculate the similarity merely based on information of the items with common scores. Therefore, the similarity between two users can be calculated given the type of the item, the reason is that, there is an average preference in the same class of items.

2.6 Implementation scheme based on scoring statistical prediction

The scoring statistical prediction programming model needs to assign a job in each step. Among them, the Map reads the number from HDFS and then outputs data, sends key value pairs through Shuffle to Reduce, which takes <key, Iterator < value >> as a n input and outputs processed key values to HDFS.

Scoring statistical prediction job means reading and writing HDFS for several times, and its output and input has some correlation. There are some problems in the algorithm based on scoring statistical prediction. The realization of a business logic requires the scoring statistical prediction, and the data exchange among jobs is completed by HDFS, increasing the expenses of the Internet and disks.

3. Experiment analysis of collaborative filtering algorithm based on scoring statistical prediction

3.1 Data selection

In choosing the nearest neighbors, the similarity calculation method should take the no score on the unknown item by the nearest neighbors into consideration. Therefore, the value of mum should not be too large to prevent reduction of predicted precision because its nearest neighbor is not the nearest neighbor of the mum in the real sense.

3.2 Data source

The data sets provided by the MOUIDLENS site include 1,100,000 comments for 1,469 films by

1,024 users. And a number ranging from 1 to 5 is selected. The greater the value, the higher preference it is. Among them, each user makes comments for at least 25 films. By calculating the proportion of items without comments in the overall user comment matrix, it is learned that 1-10000 (1024*1469) is the analysis of data sparsity.

90000 comments are randomly selected as a training set, while the others are taken as a testing set. The two sets are analyses by experiments in traditional algorithm and collaborative filtering recommendation algorithm based on scoring statistical prediction.

3.3 Result analysis

The testing results shown in table 2 include around 3.8 billion records. Compared the traditional algorithm with the collaborative filtering recommendation algorithm, it can be seen that the running efficiency and costs of traditional algorithm are obviously reduced. Among them, the scoring statistical prediction model cuts 70% of reading and writing work of HDFS, and the repeated data reading of cache, which will reduce the running time as well as the costs. While the decrease of resource scheduling can improve the running efficiency. By comparing the traditional algorithm with the collaborative filtering algorithm based on scoring statistical prediction, the running time is cut by 50%, but the cost is increased by 25%. It can be seen that the latter will effectively reduce the running time, but not in a linear manner. In a word, all items that have not received comments will be scored, and those with the highest scores will be chosen to serve as the recommendation results to repot to customers.

Table 2. Testing results

	Running time (s)	Precision (%)	Cost (%)
Traditional algorithm	65	25.4%	100
Collaborative filtering algorithm	33	67.1%	125

Therefore, the collaborative filtering algorithm based on scoring statistical prediction is able to modify and analyze similarity measure of customers when the data are sparse. In this case, it is impossible to find similar users by calculating the scoring matrix. Hence, in choosing the nearest neighbor, the scoring prediction of users based on similarity by replacing the traditional users with the first K users with similar scores has high prediction precision.

4. Conclusions

In face of complicated data processing tasks, traditional scoring statistical prediction has serious problems in performance. While most collaborative filtering recommendation algorithms have complex processing logic, which can reduce the running time and computing costs drastically in iterative computing and memory computing. In addition, it will make further improvements and modification in resource utilization, stability as well as usability so as to provide more favorable support for business and recommend more precise information for users.

References

[1] Zhou Chaojin, Wang Yuzhen. Research on the personalized recommendation of agricultural products based on improved collaborative filtering [J]. Journal of shaoyang college (natural science edition), 2014, 14(06):23-31.

[2] Tai Leilei, Tao Shiqi. A traditional Chinese medicine health care information push algorithm based on user recognition feature model [J]. Operation and management, 2016, 26(12):183-188.

[3] Wen Zhankao, Yi Xiushuang, Tian Shenshen, Li Jie, Wang Xingwei. A collaborative filtering algorithm based on the low order approximation and neighbor model of boundary matrix [J]. Computer application, 2011, 37(12):3472-3476+3486.

[4] Guo Lei, Zhang Kun, Chen Hongyan, Yan Xia. Hybrid collaborative filtering algorithm based on similarity quality [J]. Computer and digital engineering, 2015, 45(11):2099-2104.

[5] Hou Jingru, Wu Sheng, Li Yingna. Research on parallel ALS collaborative filtering algorithm based on Spark [J]. Computer and digital engineering, 2015, 45(11):2197-2201.