Research on Commodity Mixed Recommendation Algorithm

Chang Hao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 271203550@qq.com

Yang Shengquan
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 1355593006@qq.com

***Abstract*—With the advent of the era of big data, our lives generate huge amounts of data every day, and the field of e-commerce is no exception. It is particularly important to analyze these data and recommend products. It is reported that through the recommendation algorithm, Amazon has increased its sales by about 30%. Among the recommended algorithms, the collaborative filtering algorithm is currently relatively mature and has achieved very good results in various fields. But the traditional collaborative filtering algorithm is too rough when calculating the similarity and prediction score, and the efficiency is very low. We combine the traditional collaborative filtering algorithm with the decision tree algorithm, and improve the traditional recommendation algorithm, create a collaborative filtering decision tree algorithm to recommend products, and run the new collaborative filtering decision tree algorithm on the Hadoop platform on. Experiments show that the improved algorithm makes the accuracy of recommendation significantly improved.**

Keywords-E-Commerce; Recommendation Algorithm; Decision Tree; Collaborative Filtering

I. INTRODUCTION

With the development of science and technology, Internet technology has also rapidly developed and popularized, so that the data on the network is growing at the level of PB every day, bringing a lot of information resources to users and greatly enriching people's daily lives. However, the problem of rapid expansion of a large number of information resources has also emerged. "Information overload" is also a problem that Internet users are facing. "Information overload" refers to the difficulty for Internet users to accurately and quickly locate the information they need from massive data [1]. The generation of recommendation system has greatly eased the problem of "information overload". The recommendation system is automatic and intelligent to recommend items for users, and it will dynamically adjust the recommended item types according to the changes of user behavior, which truly avoids the "information overload" problem.

Faced with such a huge amount of data, it is necessary to adopt a big data model for analysis. Compared with the traditional data model using random analysis (sampling survey), the big data model analyzes all data and has the characteristics of 4V, Namely Large Volume, High Speed, Variety, Value. Collaborative filtering algorithm is one of the most concise and practical recommendation algorithms. If you use traditional data model for sampling survey, it will inevitably aggravate the sparsity problem of the algorithm itself, so it is of great significance to design a big data model based on collaborative filtering. And necessary. If you want to process big data, a single computer cannot be realized, so the application of distributed architecture is particularly important. So the algorithm model is run under the Hadoop distributed framework. MapReduce is a distributed computing framework under Hadoop [2]. It uses the "divide and conquer" idea to decompose complex tasks or data into several simple tasks for parallel processing. Afterwards, it performs global summarization, which greatly improves the efficiency of the algorithm. This article mainly studies the distributed recommendation algorithm under the Hadoop platform. The recommendation algorithm combines the decision tree and the collaborative filtering algorithm, and improves the traditional collaborative filtering algorithm to improve the timeliness of recommendation.

II. INTRODUCTION TO RELATED TECHNOLOGIES

A. Introduction to the traditional collaborative filtering algorithm

Collaborative filtering algorithm is the most successful information filtering algorithm used in the

DOI: 10.21307/ijanmc-2020-021

current recommendation system. The main method is to extract the historical behaviors generated by users to make recommendations. The traditional collaborative filtering algorithm is mainly divided into item-based collaborative filtering algorithm (ItemCF) and user-based collaborative filtering algorithm (UserCF) [3]. The core process of collaborative filtering algorithm is as follows: Collect user preferences, find similar users or items, and calculate recommendations, the core of which is the calculation of similarity Euclidean distance similarity method (Formula 1), Pearson correlation coefficient similarity method (Formula 2), Salton similarity method (Formula 3) and Cosine similarity method (Formula 4) are several common similarities Calculation method [4].

$$O(x,y) = \sqrt{(\sum(x_i - y_i)^2)} \quad (1)$$

$$P(x,y) = \frac{n\sum x_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \quad (2)$$

$$S(x,y) = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)|} \cdot \sqrt{|N(v)|}} \quad (3)$$

$$COS(x,y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \quad (4)$$

B. Introduction to Decision Tree

It is a typical classification algorithm. It first processes data, uses inductive algorithms to generate readable rules and decision trees, and then uses decisions to analyze new data. After getting the recommended products in the previous step, extract the features [5]. A data set will extract many features, but which feature is selected as the root node? Which feature is selected as the optimal solution? At this time, we need to introduce a new measure-entropy. Entropy refers to the uncertainty of random variables. The algorithm for calculating their entropy values is shown in Formula 5:

$$H(x) = -\sum pi * \log pi, i = 1,2, \dots, n \quad (5)$$

Where pi represents the probability of each feature, it can be seen from the formula that when the probability is greater and the purity is greater, the entropy value will be smaller, and the smaller the entropy value, the more stable the feature. There are three selection criteria for features: information gain (ID3), information gain rate (C4.5) and Gini index (CART). The development of decision trees is getting faster and faster, and many excellent algorithms have been derived, such as GDBT (Gradient Boosting Decision Tree), RF (Random Forest random forest), etc, the decision trees have considerable advantages in terms of accuracy improvement, they are used frequently in various games, and the effect is very good. At some times, the neural network that has been painstakingly built is not as accurate as the relatively simple random structure. The forest is high, so this article will use random forest to optimize the algorithm.

III. RECOMMENDED ALGORITHM DESIGN

The collaborative filtering algorithm is the most widely used algorithm in the recommendation system. Because of its versatility of the model, it does not require too much expertise in the corresponding data field. The engineering implementation is relatively simple, and the effect is also good. It is widely praised by all walks of life. But collaborative filtering has its own problems, such as "cold start" and "sparseness" has always been a problem of collaborative filtering itself. Therefore, in the context of today's big data era, collaborative filtering may not be suitable for direct recommendation algorithms. To solve this problem, this paper designs a hybrid recommendation algorithm.

A. Random Forest random forest

Because a single decision tree has obvious drawbacks in the recommendation system, a random forest composed of multiple decision trees can effectively solve the problems of a single decision tree. In essence, the decision tree is a special tree because it contains many judgment nodes, and the model of the random forest is composed of multiple decision trees[6].

The main idea of the random forest algorithm is to use some single classifiers to form a large classifier, which mainly includes the following three steps:

Step 1: Randomly sample multiple decision trees that need to be generated. As many decision trees as needed, as many training subsets as there should be. This process involves a statistical sampling method, which is to extract several training subsets from the original training set. The sampling method to be used in this thesis is a sampling method based on Bagging thought. When the training set is extracted, it can be sampled repeatedly to ensure that the chance of the sample being selected is random and the probability is equal.

Step 2: how to build the decision tree. The decision trees constructed from several training subsets selected in the first step are the main elements that constitute a random forest. The construction of these trees is not restricted by any factors, and does not do any pruning operations on the trees. When constructing a decision tree, not all attributes in the data set are selected as indicators for calculation, but are divided into several randomly selected "optimal" feature attributes, Then decompose according to the eigenvalue of k<K. In the decision tree, the C4.5 algorithm is used as the splitting algorithm for attribute selection, and the information gain rate algorithm is given. According to the randomness peculiar to the random forest, first select k attributes as the features of the decision tree. These features all act as classifiers. From the calculation of the training set, we can know the classification standard $h(x,\theta) \in (0,1), x \in R^N$ is a randomly selected training sample. $\theta = (\alpha,\ \varphi)$ represents the parameter of this node, φ represents the matrix, α represents the filtering function, and the surface style of the node feature is determined by α, The Formula 6 represents the calculation of the nonlinear plane, and the calculation formula of the linear plane is Formula 7:

$$h(x,\theta) = \delta(\alpha^T(x)\varphi > 0) \quad (6)$$

$$h(x,\theta) = \delta(\alpha^T(x)\varphi\alpha(x) > 0) \quad (7)$$

Use a recursive method to operate on the data set until the data on a node has all belonged to the same type of feature or the number of data sets on the node has reached the threshold set in advance, then this node will stop continuing to classify, Converted to a leaf node. If the above requirements are not met, the node will continue to randomly search for feature attributes for classification.

Step 3: the formation of the forest. After repeating the first step and the second step several times, the resulting trees can be used to build random forests. First of all, according to the function of these trees, you can classify the training set, integrate the results of the data set processed by the decision tree, and vote. The final output of the algorithm is the result of the classification with the most votes.

B. AHP model

AHP (Analytic Hierarchy Process) is a method for making decisions based on the weight of layers. This method through further exploration and analysis of the root of more complex problems and their influencing factors, further proposed a qualitative method to quantify the problem, so as to provide more detailed quantitative information for decision-making. In this research, we adopt the method of analysing the weight of influencing factors in the analytic hierarchy process, and give corresponding weights to different operation behaviors, and then determine the similarity between different users and the similarity between different brands according to the obtained weights degree. The following is the specific process of calculating the weights:

Step 1: decision-level analysis. First of all, through in-depth study of the problem, in the research process, the factors that have related relationships are analyzed and compared with each other, and each factor is arranged in layers to form a multi-layered hierarchical structure model. Through the analysis, we can know that the following four factors have the most impact on the user's future purchases: the number of user clicks on the product, whether the user has added the product to the shopping cart, whether the user has collected the product, and whether the user has purchased the product. Formed a structural model as shown below:

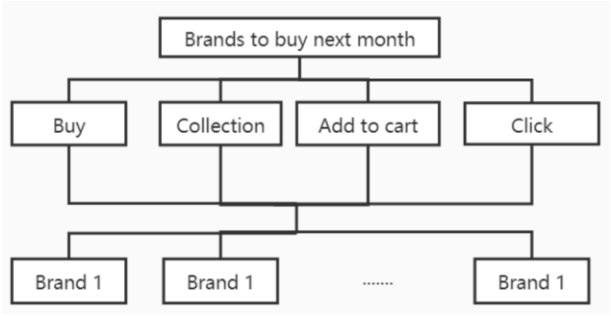

Figure 1. Hierarchical model diagram

Step 2: Construct the judgment matrix. First of all, it is necessary to compare all the influencing factors with each other. In the measurement, the introduction of relative scale is used to reduce the difficulty of comparing two different factors with each other, which further improves the accuracy.

If you want to compare the influence of n elements $A_1, A_2, \ldots, A_n$ on the same goal, you will obtain two factors A_i and A_j each time. c_{ij} represents the ratio of the influence level of A_i and A_j on the goal. All the comparison results are also can be written as a comparison matrix:

$$C = (c_{ij})n * n,\ \ c_{ij} > 0,\ \ c_{ij} = \frac{1}{c_{ij}} \quad (8)$$

The larger the value of c_{ij}, the higher the importance of A_i relative to A_j. In general, the differences between these factors are presented on a scale of 1-9.

Step 3: Solve and test. The elements of W are the ranking weights of the relative importance of the factors at the same level to the factors at the previous level. This process is called hierarchical single sorting. Then, whether the hierarchical single sorting can be confirmed requires a consistency test. The so-called consistency test is Refers to the allowed range of inconsistency determined by the pairwise comparison matrix. The consistency index can be defined as: $CI = \frac{\lambda_{max}-n}{n-1}$. Different CI values represent different meanings, the greater the CI, the more serious and obvious the inconsistency, the consistency ratio is defined as $CR = \frac{CI}{RI}$. RI is defined as a random consistency indicator, its value is shown in the table below:

TABLE I. RI VALUE TABLE

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

C. Improved collaborative filtering algorithm

Because Random Forest also has its own drawbacks, it can only recommend brands that have been in contact with users, and those brands that have not been in contact with users, Random Forest will not recommend it. Even the brand may meet the needs of users. At this time, AHP improved collaborative filtering algorithm is proposed to solve this problem [7]. Improve through the following steps.

Step 1: use the AHP model to find the weight of user behavior. Because when calculating user similarity and brand similarity, collaborative filtering recommendation algorithm cannot treat all user interactions equally. Therefore, the AHP model is used to assign values to the behaviors of users. With this set of weights, the similarity between the user's image and the brand's image can be calculated, which solves the drawback of collaborative filtering recommendation. This set of weights can be trained using the AHP analysis model introduced in the previous section.

Step 2: The user's rating data for the brand. The operation of scoring products on the e-commerce website, and the size of the rating also represents the user's love for the brand, Therefore, before calculating the user's similarity and brand's similarity, you first need to calculate the user's rating value for the brand. You can obtain the user's rating value for the brand by calculating the user operation type and frequency. The calculation formula is as follows:

$$R_{u,i} = \sum_{c=1}^{C} Op(c)Fp(c) \tag{9}$$

Among them, u represents the user, i represents the brand, then $R_{u,i}$represents the user's rating of the brand, $Op(c)$ represents the weight of the user's operation type c, and $Fp(c)$ represents the user The frequency of operations on the brand. From this we can get the matrix that $R_{u,i}$ can be composed of: $\begin{bmatrix} R_{11} & R_{12} & \dots & R_{1n} \\ R_{21} & R_{22} & \dots & R_{2n} \\ \dots & \dots & \dots & \dots \\ R_{m1} & R_{m2} & \dots & R_{mn} \end{bmatrix}$.

This is a matrix reflecting u and i interacting with each other. Based on these, the similarity between users and brands can be calculated.

Step 3: Calculation of similarity. This is the most important link in the collaborative filtering recommendation algorithm. This article uses an attribute-based similarity algorithm. The similarity is to process the information of the user and its nearest neighbors, so the core part of the algorithm is that if the nearest neighbor is obtained, we conduct relevant research and analysis on it, and obtain the following method of the user's comprehensive similarity.

The following is the formula for the user's rating similarity $YSD1(u, i)$:

$$YSD1 = \frac{\sum_{t\in I_{u,i}}(R_{u,t}-A_u)*(R_{i,t}-A_i)}{\sqrt{\sum_{t\in I_u}(R_{u,t}-A_u)^2 * \sum_{t\in I_i}(R_{i,t}-A_i)^2}} \tag{10}$$

This formula compares the similarity of users u and i. $R_{u,t}$ and$R_{i,t}$each represent the rating value of users u and i on brand t, and the set of brands that user u has rated respectively represent For I_u. Similarly, all brands rated by user i. are denoted as I_i, and the intersection of those users who have shared a rating is $I_{u,i}$; the average value of user u's rating in I_u is set to A_u, similarly, in I_i The average value of user i score is set to A_i.

The following formula represents the user's feature similarity YSD2 algorithm:

$$D(u,i) = \sqrt{\sum_{k=1}^{n}(u_k - i_k)^2} \quad (11)$$

$$YSD2(u,i) = \frac{1}{1+D(u,i)} \quad (12)$$

Formula 12 shows the weighted Euclidean Metric of users u and i, that is, the Euclidean distance, n represents the feature dimension of the user, and the k-th eigenvalues of users u and i will be represented by u_k and i_k, respectively. YSD2 illustrates the calculated similarity of features of users u and i.

Based on the above calculation formula to obtain the user's similarity score YSD1 for the product and the user characteristic similarity YSD2 , the following formula is used to calculate the user's comprehensive similarity YSD:

$$YSD(u,i) = w_1 * YSD1(u,i) + (1 - w_1) * YSD2(u,1) \quad (13)$$

In these indicators, w_1 is the combined weight of the user's comprehensive similarity. The actual value of the combined weight of the user's comprehensive similarity is determined by the degree of influence of the score similarity and the feature similarity on the user's comprehensive similarity.

The calculation of brand similarity is the same as the calculation of user similarity. You only need to change the parameters, which will not be described in detail here.

Step 4: the selection of the nearest neighbor. In order to achieve the purpose of accurate recommendation, it is necessary to accurately target the other neighbor users that match the user's interests, so the selection of the nearest neighbor is very important. Select the user's nearest neighbor and the brand's nearest neighbor, this article uses the Top-N method. The first step of this method is to calculate the similarity between other users, brands and target users, brands, and then sort the calculated similarity values.

Step 5: Generate recommendations. Using the method described in step 3, we can obtain the nearest neighbor set NU[u] of the target user, and then we recommend the user according to the formula listed below:

$$PU(t,u) = A_u + \frac{\sum_{i=1}^{c}(R_{i,t} - A_i) * YSD(u,i)}{\sum_{i=1}^{c} YSD(u,i)} \quad (14)$$

A_u is the average score of user u for all brands in the dataset, the value of c is the number of users in the nearest neighbor of user u., $R_{i,t}$ is the nearest neighbor user i. of user u, For the rating made by the brand t, $YSD(u,i)$ represents the user's overall similarity. PU means that the recommendation degree of this brand is recommended to the user u based on the result of the user recommendation.

The brand-based recommendation idea is similar to the above, so I won't elaborate more.

D. Fusion of the two algorithms

After introducing the basic characteristics of the random forest recommendation algorithm and the collaborative filtering recommendation algorithm, the parameters and the calculation process required, we use the characteristics of each other. For example, the random forest model can recommend brands that users have interacted with before, while the collaborative filtering algorithm recommends brands that have not interacted with users. Therefore, if the two models can be perfectly combined, the common advantages of the two algorithms can be combined, which plays a role in promoting strengths and avoiding weaknesses. All the information required for user recommendation can be included. Naturally, better recommendation results will be obtained. The model the accuracy is also more improved, as shown in detail in Figure 2 describes the fusion strategy of the two models:

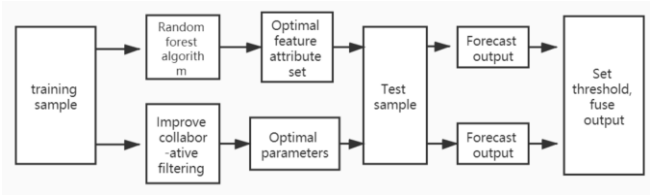

Figure 2. Schematic diagram of algorithm fusion process

IV. EXPERIMENT AND RESULT ANALYSIS

A. Experimental data and experimental environment

In order to verify the efficiency of the improved algorithm model, we selected an e-commerce company's internal data set for testing, and analyzed and evaluated its performance. The data set contains all the behaviors of about 100,000 random users with behaviors within one week of the e-commerce company. Among them, user behaviors include clicks, purchases, favorites, and purchases. The data set contains users 99799, the number of goods 416202, the number of all actions 10015080, such a huge data set, if

you use traditional stand-alone operations, the time consumption is immeasurable, so using the Hadoop distributed system in a big data environment to perform distributed calculations on its data greatly improves the efficiency of the operation.

The experimental operating environment is a Hadoop cluster. The cluster has one Master node and three Slave nodes, and all machines in the cluster have the same configuration. The cluster installation is configured under the CentOS-6.7 operating system, and the JDK1.8 environment is configured for both CentOS and Windows, and the code is compiled into the IDEA compiler on the Windows side.

B. Algorithm evaluation index

1) Recall rate

The number of items in the recommended list calculated by the algorithm is what consumers like, which is the recall rate of the algorithm. Formula 15 describes how to calculate the recall rate:

$$R(L_u) = \frac{L_u \cap B_u}{B_u} \quad (15)$$

Among them, the item that user u likes is L_u, and the recommendation algorithm lists the product recommended by the consumer as B_u.

2) Accuracy

The accuracy of the algorithm tests the ratio of the items in the recommended list given by the system to the items that consumers actually like. Formula 16 describes how to calculate the accuracy:

$$P(L_u) = \frac{L_u \cap B_u}{L_u} \quad (16)$$

Among them, the item that user u likes is L_u, and the recommendation algorithm is denoted as B_u by the consumer's recommended product list.

3) F measure

As shown in Formula 17:

$$F = \frac{2*P*R}{P+R} \quad (17)$$

Since there is a negative correlation between the accuracy rate and the recall rate, it is necessary to fit the F measure, and the F score will prevail. As can be seen from the Formula 18, the prediction result hopes to cover more users and brands while ensuring accuracy.

C. Random forest model experiment results

There are two parameters that can be controlled by the random forest model, one is the number of random forest decision trees, and the other is the number of feature attributes that are randomly extracted to build the decision tree. The number of decision trees in the random forest model is K_{RF}, and the selected parameters are K_{RF}=100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, and the following experimental results can be obtained:

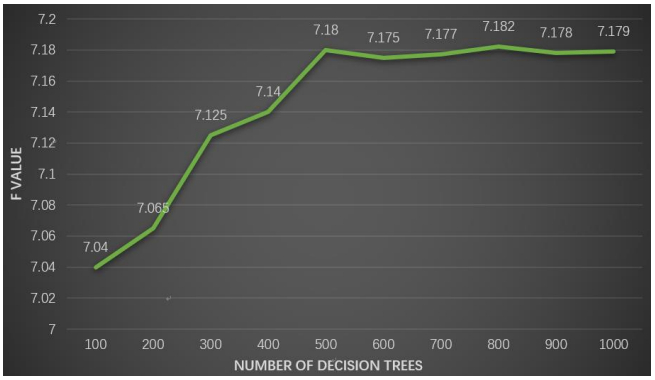

Figure 3. The relationship between decision tree and result in random forest model

It can be seen from the experimental result graph that the F value rises with the increase of the decision tree. After rising to 500, it basically tends to be gentle, so K_{RF} is taken as 500.

Also for the eigenvalues of random forests, not as many as possible, in order to ensure the stability of the model, through experiments, the following results are obtained:

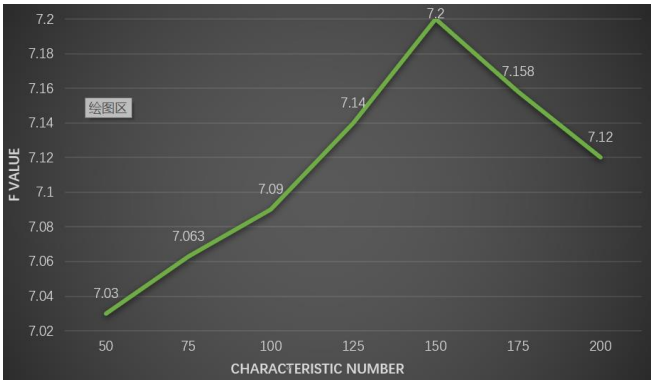

Figure 4. The relationship between the number of features and the results in the random forest model

It can be seen from the above experiment that when the number of decision trees is 500 and the number of features is 150, the random forest model has the best effect, that is, the F value is the highest.

The final experimental results of the random forest model are shown in the table below.

TABLE II. RANDOM FOREST MODEL FINAL EXPERIMENTAL RESULTS

Accuracy	Recall rate	F measure
7.19	7.21	7.20

D. Experimental results of the hybrid algorithm

The weight of user behavior in the improved collaborative filtering algorithm can be calculated according to AHP, as shown in the following table：

TABLE III. USER BEHAVIOR WEIGHT

Interaction type	weight
Click	0.08
Collection	0.12
Add to cart	0.30
Buy	0.50

The number n of brands recommended by the collaborative filtering algorithm is used as the parameter for the experiment. We take n = 600, 800, 1000, 1200, 1400, and 1600 to conduct the experiment. The experimental results are shown in the figure.

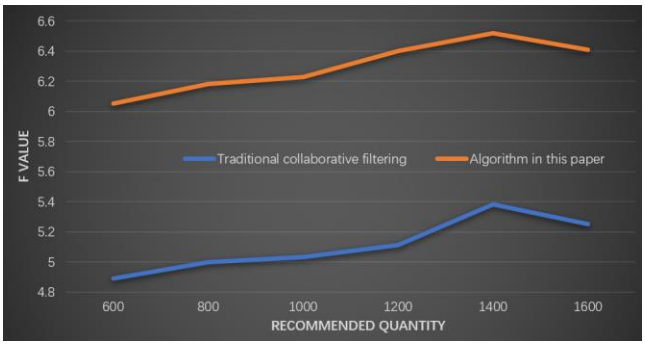

Figure 5. The relationship between the recommended number of collaborative filtering algorithms and the result

It can be seen from the experimental results that with the increase in the number of recommendations, the algorithm's recommendation performance is improving. When the number of brand recommendations is close to 1400, the recommendation effect is the best.

The final experimental results of the fusion of the final collaborative filtering recommendation algorithm and the random forest algorithm are shown in the following table:

TABLE IV. EXPERIMENTAL RESULTS OF FUSION ALGORITHM

Accuracy	Recall rate	F measure
7.33	7.42	7.35

Due to the large number of data sets in this experiment and the time-consuming operation in a single machine, the above operations are performed under the distributed system Hadoop. In order to compare the advantages of Hadoop in data processing speed, the collaborative filtering and the design The algorithm and the processing time of the algorithm designed in this paper are compared in the Hadoop environment, as shown below:

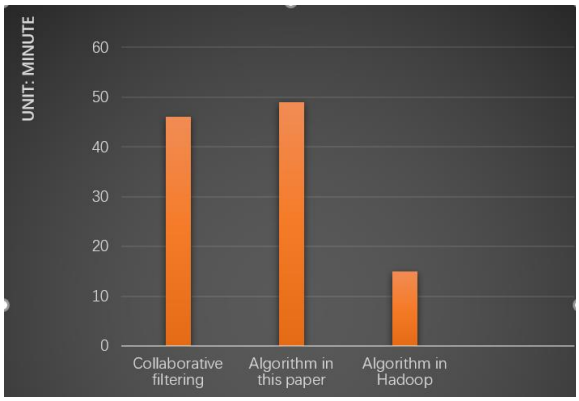

Figure 6. Time comparison chart

Because the algorithm in this paper is more accurate than the traditional collaborative filtering algorithm and the calculation is relatively complicated, the time efficiency is slightly insufficient. However, if it is run in a Hadoop distributed cluster environment, the time efficiency is effectively improved by nearly 3 times. It can be seen that the current big data In a large environment, it is necessary to use Hadoop distributed clusters to process data.

V. CONCLUSION

This paper attempts a model that uses different prediction and recommendation methods for prediction and recommendation, namely the random forest model, and gives a detailed introduction and further analysis of this model. This article also gives a detailed

introduction to the basic principles of the traditional filtering algorithm of collaborative filtering, and thoroughly analyzes the advantages and disadvantages of the algorithm. For example, this traditional collaborative filtering algorithm lacks the ability to calculate the brand score of users. Personalized investigation, treat all user behavior as the same. In view of the limitation of this traditional algorithm, this paper made some necessary improvements, and proposed the optimization of collaborative filtering similarity based on the weight of AHP. In addition, from two perspectives, user interaction and brand interaction, this article randomly integrates the collaborative filtering model with the random forest model. This will result in more accurate recommendation results, and the recall rate will naturally increase. Finally, the data is analyzed through real data cases to obtain reliable experimental results. The results show that the combination of this analytic hierarchy process and collaborative filtering algorithm makes the recommendation performance better than a single collaborative filtering algorithm. And after being fused with the random forest model, compared with the single random forest algorithm or collaborative filtering algorithm, the performance has been greatly improved.

REFERENCES

[1] Lu Xiaocui. The application of big data analysis technology in cross-border e-commerce[J]. Electronic Technology and Software Engineering, 2020 (01): 141-142.

[2] Tian Bin. Big data machine learning under the framework of distributed computing[J]. Electronic Technology and Software Engineering, 2019(20):168-169.

[3] Yang Wu, Tang Rui, Lu Ling. News recommendation method combining content-based recommendation and collaborative filtering[J]. Computer Applications, 2016, 36(02):414-418.

[4] Yang Hailong. Power recommendation system based on item-based collaborative filtering algorithm[D]. Lanzhou Jiaotong University, 2019.

[5] Sheng Wenshun, Sun Yanwen. An improved ID3 decision algorithm and its application[J]. Computer and Digital Engineering, 2019, 47(12):2943-2945+3094.

[6] Wang Jingna. Research on used car valuation model based on random forest algorithm[D]. Beijing Jiaotong University, 2019.

[7] Cui Yan, Qi Wei, Pang Hailong, Zhao Hui. A recommendation algorithm combining collaborative filtering and XGBoost[J]. Computer Application Research, 2020, 37(01):62-65.