Distributed Recommendation Algorithm Based on Fuzzy Clustering

Jiali zhang1,*, Haohua Qing2
1Department of Computer Science and Engineering, Guangzhou College of Technology and Business, Guangzhou, 510850, Guangdong, China
2Department of Economic And Trade, Guangzhou College of Technology and Business, Guangzhou, 510850, Guangdong, China

*Corresponding author E-mail:155128373@sasu.edu.cn

Abstract. In recent years, the Internet is sneaking into every corner of users' lives at a speed visible to the naked eye. Information has shown a geometric growth with the development of the Internet. This article mainly studies the application of distributed recommendation algorithm based on fuzzy clustering. In this paper, the data set used is divided into training set, test set and validation set in a random manner. The ratio of the three number sets is 7:1:2. This paper randomly selects 70% of consumers' historical data as the training data set, and the last 10% as the test set. The USCensus1990raw data set is divided into three groups, numbered 1, 2, and 3, and the number of samples is 400000, 600000 and 800000 respectively. Under the Hadoop platform, the MrDPCND-FCM algorithm is used to cluster the three sets of data to be tested, the number of nodes is increased from 2 to 5, and the time taken to cluster different data sets to be tested under different numbers of nodes is recorded. Therefore, offline testing uses items that the user has already evaluated. If the user has not scored the item, it cannot be evaluated. When the number of clusters is 60, the data sparsity drops from 92% to 51%. When the number of clusters reaches 5, the sparsity drops to around 3%. The algorithm in this paper solves the problem of large-scale data to a certain extent, and improves the efficiency and accuracy.

Keywords: Fuzzy Clustering, Distributed Recommendation Algorithm, Data Sparsity, Dynamic Adjustment

1. Introduction

With the vigorous development of information technology, people have entered the information age, web data is showing an exponential growth situation, and overloaded Internet data makes it more and more difficult for people to obtain information. Followed by the emergence of search websites, this
type of website sorts out some commonly used network resources to facilitate people's search. Cluster analysis has also been widely used in the fields of climatology, psychology and medicine. Using cluster analysis to discover the patterns of the atmosphere and the ocean can better understand the terrestrial climate and forecast the climate.

From the generation to development of recommendation system, in terms of meeting the user's recommendation needs, from individual user recommendation to group user recommendation, from single information recommendation to multiple information recommendation, and the continuous optimization of various recommendation models, the accuracy of recommendation is constantly improving, which is increasingly meeting people's information acquisition needs [1-2]. Today, with the rapid development of Internet of things, personalized recommendation system has become one of the most important means of enterprise promotion and publicity [3]. From the perspective of users, a mature recommendation system can provide users with better experience and even more convenient life [4]. For enterprises, high-quality recommendation system can reduce the loss of users, thus creating greater economic benefits [5-6]. Different from the classification algorithm, the clustering algorithm does not rely on the pre-set categories or the training data set that has labeled the samples, so it needs the clustering algorithm to label the categories in the way of autonomous learning [7-8]. Therefore, we can use fuzzy clustering to fuse the implicit feature information obtained from implicit semantic decomposition with the explicit information of items or users, so as to improve the accuracy of the recommendation algorithm [9-10].

Under the fierce competition of e-commerce, recommendation algorithm can bring great benefits to the rapid development of e-commerce industry. The biggest advantage of recommendation system is that it can collect user's interest information and make personalized recommendation for users according to their different preferences. Moreover, the recommendation information is dynamically updated, that is to say, with the passage of time, the user's interest is gradually changing, and the recommendation result of recommendation system will also change. Therefore, the recommender system greatly improves the user experience of the website and facilitates users to query resource information.

2. Distributed Recommendation Algorithm Based on Fuzzy Clustering

2.1. Fuzzy Clustering Algorithm

The clustering process is not guided by any other prior knowledge. All objects in each cluster have high similarity, but the similarity between different clusters is very low. The similarity calculation formula of the consumer score feature vector can use the calculation formula with commodity penalty, the formula is as follows:

$$w_{uv} = \frac{1}{\log(1+N(i))}$$

$$\sum_{i \in N(u) \cap N(v)} \sqrt{|N(u)| \times |N(v)|}$$

(1)
In the formula, \( W_{uv} \) is the similarity of preferences between user \( u \) and user \( v \).

The formula for calculating the preference of the consumer \( u \) to be recommended for the recommended product \( i \) is as follows:

\[
p(u, i) = \sum_{f=1}^{F} W_{uf} r_{fi} p
\]

In the formula, \( p(u, i) \) represents user \( u \)'s preference for item \( i \).

The fuzzy clustering model is as follows:

\[
\begin{align*}
\min_{U,V} J_{K-\text{means}}(U,V) &= \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij} d_{ij}^2 \\
\text{s.t.} u_{ij} &\in \{0,1\}, \sum_{i=1}^{c} u_{ij} = 1.0, \sum_{i=1}^{c} u_{ij} < n
\end{align*}
\]

The quality of a clustering algorithm can be judged by whether it can find some unknowledged classification rules in the data. Different clustering algorithms may be used to cluster the same data set to obtain different clustering results. Cluster analysis is a key research direction in the field of data mining, and the research on it is also a very challenging task. Through the parameter server, the row and column of the parameters being used are explicitly locked to ensure that the problem of lagging gradient will not occur, so that the convergence of the loss function can get the same effect as the single-machine serial training. However, parameter locking will cause a large number of computing nodes to wait for synchronization, and in extreme cases the efficiency will even be lower than that of single-machine serial training.

2.2. Distributed Recommendation Algorithm

Recommendation algorithms generally analyze user historical behavior records, and its performance directly affects the evaluation indicators and the quality of the recommendation results. Many algorithms will measure the similarity of users or items when analyzing user behaviors, and produce recommendation results based on these similarities. User behavior record and similarity measurement are very important for recommendation algorithms. Moreover, peer-to-peer networks are often dynamic, and nodes can join or leave the network in real time. First, based on the collected raw data, extract the user's rating information on the item, analyze the rating information, build a user or item similarity neighbor set, and finally sort the similarity to generate a personalized recommendation result for the user.

3. Fuzzy Clustering Model Experiment

3.1. Experimental Environment
Hadoop cluster configuration information is shown in Table 1. The cluster used in the experiment in this article consists of three machines, one of which is the Master, which is configured as the Name Node node, and the other two are configured by the Slave.

Table 1. Hadoop cluster configuration information

| Node name | IP address  |
|-----------|-------------|
| weekend01 | 192.168.78.66 |
| weekend02 | 192.168.78.77 |
| weekend03 | 192.168.78.88 |

3.2. Evaluation Index

In this paper, the data set used is divided into training set, test set and validation set in a random manner. The ratio of the three number sets is 7:1:2. This paper randomly selects 70% of consumers' historical data as the training data set, and the last 10% as the test set. The US Census1990raw data set is divided into three groups, numbered 1, 2, and 3, and the number of samples is 400000, 600000 and 800000 respectively. Under the Hadoop platform, the MrDPND-FCM algorithm is used to cluster the three sets of data to be tested, the number of nodes is increased from 2 to 5, and the time taken to cluster different data sets to be tested under different numbers of nodes is recorded.

4. Discussion

4.1. Recommended Accuracy Analysis

The comparison of FCM algorithm iteration times is shown in Table 2. From the table, we can see that as the number of clusters increases, the iteration time of the traditional FCM algorithm increases significantly. However, SWFCM has improved due to the decrease of input samples and the increase of iteration time. Although the number of iterations has not increased rapidly, there have been significant changes. However, the improved algorithm in this paper has only small changes, so it is proved that the algorithm improves the robustness of the traditional algorithm to a certain extent. Under different number of clusters, the sparsity of the data is different. The fewer the number of clusters, the fewer the number of non-zeros in the data, and the smaller the data sparsity. When the number of clusters is 60, the data sparsity drops from 92% to 51%. When the number of clusters reaches 5, the sparsity drops to around 3%. Therefore, the problem of data sparsity can be solved very well by fuzzy mean clustering. But on the other hand, it can also be seen that the growth of the acceleration ratio and the growth of the number of cluster nodes do not show a linear growth relationship. This is because the communication and scheduling between the nodes are also consumed to a certain extent while the cluster nodes are increased. Resources, and make the acceleration ratio increase more slowly than the number of nodes. For data sets of different sizes, the larger the data set, the more obvious the increase in the speedup ratio. This is because as the amount of data increases, the proportion of time occupied by each node in the algorithm process becomes larger. The proportion of time occupied by the mutual communication between nodes is reduced, making the acceleration ratio change more obvious.
Table 2. Comparison of iteration times of FCM algorithm

| Number of categories | Category 6 | Category 8 | Category 10 | Category 15 |
|----------------------|------------|------------|-------------|-------------|
| FCM                  | 55         | 91         | 208         | 241         |
| SWFCM                | 30         | 36         | 42          | 64          |
| NSWFCM               | 21         | 20         | 23          | 27          |

4.2. Algorithm Performance Analysis

The accuracy of the algorithm is shown in Figure 1. The dimension of the preprocessed data is reduced from 41 dimensions of the original data to 7 dimensions. The data structure of RDD becomes simple, the memory storage is relatively small, and the algorithm iteration of spark platform is relatively fast, which is consistent with the theoretical time-consuming trend prediction. This is because: the traditional collaborative filtering needle directly constructs the user item score matrix on the original data set, then calculates the similarity, sorts and filters out the most similar users, the algorithm does not consider the influence of matrix sparsity and user background information on the recommendation results, and the similarity calculation of all data increases the amount of calculation and time-consuming. When the data set is small, there is almost no obvious difference between the single machine and the distributed system. However, with the growth of the data set, the time-consuming of the distributed system based on MapReduce is far less than that of the single machine. When the data set size is more than 1 million, the time-consuming of the distributed system with three nodes is 912 seconds less than that of the single node system. The algorithm achieves the expected effect, and the recommendation efficiency is significantly improved compared with single machine in massive data.

![Figure 1. The accuracy of the algorithm](image)

The average running time of each algorithm is shown in Figure 2. The FCM algorithm using Euclidean distance takes the shortest time, but because Euclidean distance can not effectively deal with the time drift characteristics, it only takes this time as a reference. After increasing the weight of distance, the point far away from the selected initial center is preferred, and even outliers may be selected, so that the really better initial center is ignored in the selection process, which affects the accuracy and recall of the algorithm and reduces the accuracy of the algorithm. When the K value is 5, the items that participate in the target scoring prediction have a high similarity with the target items,
which has the greatest impact on the scoring prediction. However, due to the small number, it can not fully reflect the status and relationship of the target items in the whole item space. Maybe there are some other potential characteristics of the object, which can not be fully expressed by similar objects in such a small nearest neighbor subset. It will only over strengthen some characteristics of the target object and limit it to the subset of some similar objects, which is very unfavorable for the accuracy and diversity of prediction. From the average execution time RT, it can be seen that the average execution time of the proposed algorithm and PFCM algorithm is much longer than that of the traditional FCM algorithm. The reason is that both the proposed algorithm and PFCM algorithm introduce PSO algorithm to replace the iterative process of the traditional FCM algorithm, so their average execution time is greater than that of the FCM algorithm. The average execution time of this algorithm is slightly less than that of PFCM algorithm. The reason is that this algorithm selects the cluster center according to the sample number, while PFCM algorithm selects each dimension of the sample, so the average execution time of this algorithm is less than that of PFCM algorithm.

![Figure 2. Average running time of each algorithm](image)

5. Conclusion

In recent years, thanks to the vigorous progress of information technology, the increasing scale of data, the increasing number of users and the sparse user related evaluation data, the accuracy of recommendation system is inevitably affected in this environment.

Recommender system can not only solve the problem of excessive information and show the information users need, but also has a wide range of applications in the field of e-commerce and consulting. Through the distribution of data in the sliding window with variable size, the paper evaluates whether the concept drift occurs before and after the data stream.

In the current research of personalized recommendation, it is mainly divided into the defects of personalized recommendation algorithm and effective improvement; friendly user experience design of personalized recommendation system, accurate positioning of users' similar interest clusters, improving the quality of recommendation results.
References

[1] Patel D . Threshold based partial partitioning fuzzy means clustering algorithm (TPPFMCA) for pattern discovery[J]. International Journal of Information Technology, 2020, 12(1):215-222.

[2] Huawei Y I , Zaiseng N , Fuzhi Z , et al. Robust Recommendation Algorithm Based on Kernel Principal Component Analysis and Fuzzy C-means Clustering[J]. Wuhan University Journal of Natural Sciences, 2018, 23(002):111-119.

[3] Liu S , Jin S . 3-D Gravity Anomaly Inversion Based on Improved Guided Fuzzy C-Means Clustering Algorithm[J]. Pure and Applied Geophysics, 2020, 177(2):1005-1027.

[4] Yin D , Hou R , Du J , et al. SAR image change detection method based on intuitionistic fuzzy C-means clustering algorithm[J]. Journal of Intelligent and Fuzzy Systems, 2020, 38(4):3595-3604.

[5] Xiangxiao L , Honglin O , Lijuan X . Kernel-Distance-Based Intuitionistic Fuzzy c-Means Clustering Algorithm and Its Application[J]. Pattern Recognition and Image Analysis, 2019, 29(4):592-597.

[6] Di Martino F , Pedrycz W , Sessa S . Spatiotemporal extended fuzzy C-means clustering algorithm for hotspots detection and prediction[J]. Fuzzy sets and systems, 2018, 340(6):109-126.

[7] Banupriya C V , Deviaruna D . K-modes and Fuzzy C-means with modified Particle Swarm Optimization Clustering Algorithm for Epilepsy Seizure Data[J]. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING, 2019, 7(1):73-77.

[8] Gligori M V , Gligori Z M , Belji E R , et al. Long-Term Room and Pillar Mine Production Planning Based on Fuzzy 0-1 Linear Programming and Multicriteria Clustering Algorithm with Uncertainty[J]. Mathematical Problems in Engineering, 2019, 2019(4):1-26.

[9] Misbahuddin M , Ratna A A P , Sari R F . Dynamic Multi-hop Routing Protocol Based on Fuzzy-Firefly Algorithm for Data Similarity Aware Node Clustering in WSNs[J]. International journal of computers, communications & control, 2018, 13(1):99-116.

[10] Baykasolu A , Ilker Glicük, Ozsoydan F B . Improving fuzzy c-means clustering via quantum-enhanced Weighted Superposition Attraction algorithm[J]. Hacettepe University Bulletin of Natural Sciences and Engineering Series B: Mathematics and Statistics, 2019, 48(2):1-24.