Application of Travel Service Recommendation Algorithm Based on Cloud Computing

Aronna\textsuperscript{1} and Sulongaowa\textsuperscript{2,*}
Tourism College of Inner Mongolia Normal University, Hohhot, 010022, China

*Corresponding author e-mail: Sulongaowa@imnu.edu.cn

Abstract. In the context of big data, traditional collaborative filtering recommendation algorithms cannot provide users with accurate recommendation services, making the sparsity of user data become an important factor affecting the accuracy of recommendation in the complex social network environment. This paper mainly studies the design and application of travel service recommendation algorithm based on cloud computing. In this paper, the data is processed based on MapReduce parallel computing framework to improve the performance and speed of the algorithm. The Linux-equipped cluster is deployed under the Hadoop framework. The cluster feasibility test is carried out to realize the recommendation function of the algorithm, and the function of the algorithm is realized according to the actual problems. On the Hadoop big data platform, the traditional algorithm was improved and the hadoop-based collaborative filtering recommendation algorithm was implemented to further improve the recommendation rate and accuracy of the traditional recommendation filtering algorithm.

Keywords: Cloud Computing, Travel Services, Recommendation Algorithms, Hadoop Platform

1. Introduction
Tourism service recommendation is a highly recommended marketing strategy in today's tourism industry. It has been widely studied and, to some extent, it is the embodiment of wisdom. As travel data grows ever larger, storing and processing it becomes a challenge. As another way to solve the information load, recommendation engine can not only find the appropriate or potential information for users through recommendation technology, but also bring better experience to users. In order to make personalized recommendation better deal with massive data and better achieve the recommendation effect, how to improve the recommendation rate in the personalized recommendation system becomes the key point [1]. Hadoop big data analysis platform is the current data of cloud computing, the important measures to solve the problem of big data, based on the characteristics of unique Hadoop distributed file system, users using Hadoop technology platform for the construction of distributed, the handling of the work become more convenient, and can make full use of the Hadoop cluster, for the calculation of huge amounts of data to get and deal with huge amounts of data information.
With the further popularization of mobile Internet and the further development of mobile e-commerce, recommendation system has gradually developed into an important basic research method and content of E-COMMERCE recommendation system IT technology, which has attracted the attention of more and more experts and researchers. Saeed think clustering is a collection of data mining in physical or abstract objects into multiple classes of similar objects, the method of user model was generated by user behavior data, clustering techniques to cluster analysis, the user behavior, interests, similar to the user to the same class, after the clustering, can be done in search both belong to the category of users based on the recommendations from the user, not only reduces the search range, saves time, and the accuracy is high, can effectively improve the performance of the system [2]. Barman proposed to use Pearson correlation coefficient to improve the traditional Mini Batch K-means clustering distance iterative calculation method, and used the improved clustering algorithm to fill in the missing item prediction score, with significant effect [3].

The main research content of this paper is to obtain massive tourism data and study tourism service recommendation. This paper takes the hadoop-based collaborative filtering recommendation algorithm as the research direction, improves the algorithm based on the Hadoop big data platform, and makes use of the advantages of the Hadoop platform to achieve a high rate of recommendation for the personalized recommendation system to meet people's increasing demand for data information.

2. Tourism Recommendation Algorithm Based on Hadoop Platform

2.1. Hadoop Related Technologies

Hadoop is a framework for large-scale data processing, which is free to use. Distributed applications on Hadoop can also be written directly using other types of language applications, and the Hadoop ecosystem has its own core components [4].

Hadoop is accurate, high-speed, scalable, and relies on its unique data processing method to process massive data for users. The Hadoop ecosystem is composed of HDFS, MapReduce, HBase, Zookeeper, Hive, Sqoop, Flume and other components that are closely connected. The core of Hadoop design lies in HDFS and MapReduce [5].

(1) HDFS: Highly fault-tolerant and based on Linux local file system, improved information transfer rate, suitable for those large-scale data applications. HDFS is designed with the idea of separate governance in mind. The ability to divide large data sets into smaller files that can be stored on more servers in parallel makes it easy to perform a series of calculations and analyses on large data sets. HDFS can provide massive data storage services for distributed computing frameworks such as MapReduce and Spark, which fully reflects its role in the big data system.

(2) MapReduce: an important component in Hadoop parallel computing, which can realize the distributed programming model of parallel computing on many machines. It can accomplish the parallel processing of TERabyte-level data in the cluster, extract the elements we need from it and return the results. In the operation process of MapReduce, we decompose relatively large data and divide the large file into several small pieces to analyze and calculate one by one. Finally, we carry out a series of summary analysis on the calculated data, and then we can get the content we need.

(3) HBase: HBase is a highly reliable, object-oriented column, scalable and real-time read-write distributed database. HBase uses distributed file system as its own file storage system and MapReduce to further process large-scale data sets. HBase is mainly used to store loose data.

(4) Zookeeper: Hadoop distributed coordination service with high availability. Zookeeper can provide maintenance configuration information service. When many projects need to be configured, it can be extracted and stored in a certain place and monitored. When there is a change in the project information, Zookeeper will be notified and apply the newly modified information to the system, which is a good way to avoid system restart [6-7]. In addition, Zookeeper provides naming and distributed synchronization, and many Hadoop components depend on it.
(5) Hive: Hive is a huge warehouse of data, much like our everyday warehouses. Hive relies on HDFS to process structured data through SQL. Unlike databases, Hive is a tool that does not store real data. Hive's function is to turn query statements into parallel task runs.

2.2. Hadoop - Based Recommendation Algorithm for Collaborative Filtering

(1) Algorithm improvement

Based on the user interest similarity measurement method, it is believed that the similarity between users is related to the user's interest degree in a certain item. By combining the similarity of the user's interest degree in items, the user's nearest neighbor and the new predicted score are finally generated [8]. The calculation formula is as follows:

\[ \text{sim}_i(u, v) = \frac{\sum_{i=1}^{n} l_{ui} l_{vi}}{\sqrt{\sum_{i=1}^{n} l_{ui}^2} \sqrt{\sum_{i=1}^{n} l_{vi}^2}} = \frac{\sum_{i=1}^{n} N_{ui} N_{vi}}{\sqrt{\sum_{i=1}^{n} N_{ui}^2} \sqrt{\sum_{i=1}^{n} N_{vi}^2}} \]  

(1)

Interest-based similarity measurement method combines user's score and user's interest, as shown in Formula (2):

\[ \text{sim}(u, v) = (1 - \lambda) \frac{\sum_{i \in \{u \cap v\}} R_{ui} R_{vi}}{\sqrt{\sum_{i \in \{u \cap v\}} R_{ui}^2} \sqrt{\sum_{i \in \{u \cap v\}} R_{vi}^2}} + \lambda \frac{\sum_{i \in \{u \cap v\}} N_{ui} N_{vi}}{\sqrt{\sum_{i \in \{u \cap v\}} N_{ui}^2} \sqrt{\sum_{i \in \{u \cap v\}} N_{vi}^2}} \]  

(2)

After obtaining the recommendation set of items, filter and sort the items according to their rating information, and select top-n items to recommend to users.

(2) Algorithm Implementation

Based on the Hadoop big data analysis platform, this paper uses MapReduce programming framework to process data for large-scale data sets, and uses the Hadoop cluster set up to improve the running speed of the system, which can well solve the real-time requirement of the recommendation process [9]. Parallel processing is carried out based on MapReduce programming model, and the master node is divided into several sub-nodes, which cooperate with each other to complete the operation of large-scale data set. After that, the operation results of each sub-node are integrated to obtain the final result [10].

The improved collaborative filtering recommendation algorithm was programmed by MapReduce on the Hadoop distributed computing platform, and the traditional collaborative filtering recommendation algorithm was further improved based on the Hadoop big data analysis platform. The operation steps of hadoop-based collaborative filtering recommendation algorithm are as follows:

1) Read the original score data set and de-weight the data;
2) Group according to users, combine all possible items to get user rating matrix;
3) Make statistics on the items in the data set to obtain the co-occurrence matrix between the items;
4) For the convenience of future processing, the user-scoring matrix generated in the second step and the co-occurrence matrix appeared in the third step were transformed for subsequent similarity calculation;
5) Use formula (2) to calculate the improved similarity calculation;
6) Generate the nearest neighbor and obtain the user's recommendation vector matrix;
7) Filter and sort the results, and save the recommended results to the database.

In the above steps, the most critical step is to get the neighbor users with the most similar liking degree according to the calculation result of similarity. When recommending the obtained results, they should be arranged according to the recommended weight to obtain the final top-n sorting result.

3. Hadoop Cluster Construction and Experiment

3.1. Experimental Environment
This section mainly introduces the process of cluster setup by the algorithm. The algorithm is mainly built on the MapReduce distributed computing framework. Based on the free VMware Workstation 14 Player, 6 virtual machines are installed to set up Hadoop cluster for implementation.

3.2. Cluster Construction
This paper mainly runs on Hadoop cluster, so this paper uses virtual machine to set up Hadoop cluster. In addition, this paper also configured other components involved, such as Zookeeper, Hive, etc.

The setup of Windows 7 64-bit operating system is required for the experiment, and VMware Workstation 14 virtual machine software is installed in this setup. In VMware virtual machine, it can simulate the installation of many PERSONAL computers of Linux system, and connect many personal computers with NAT mode network connection. Deploy various services on various nodes in the Hadoop cluster, including HDFS, MapReduce, Zookeeper, and so on.

3.3. Experimental Data Set
Collaborative filtering is a representative recommendation method. The experimental data of this paper selects different score data of users on tourism projects as the original data. In this paper, three data sets with different sizes of 100KB, 1MB and 10MB were selected for experiments, and the Hadoop collaborative filtering recommendation algorithm proposed in this paper was implemented for running and debugging.

3.4. Experimental Evaluation Criteria
In this paper, the absolute deviation MAE and acceleration ratio S are calculated as evaluation indexes to reflect the accuracy of the algorithm. The formula is as follows:

\[ MAE = \frac{1}{n} \sum_{i=1}^{n} |p_{i} - q_{i}| \]  

4. Experimental Results of Hadoop Recommendation Algorithm

4.1. Acceleration Rates for Different Data Sets

![Figure 1. Graph of acceleration ratio](image)

As shown in Figure 1, the horizontal axis means the number of nodes in the Hadoop cluster, and the vertical axis means the value of acceleration ratio S. As can be clearly seen from the figure above, with the increase in the number of nodes in the cluster, the value of acceleration ratio for data sets of different sizes of 100KB, 1MB and 10MB also increases continuously. For the same data set,
increasing the number of nodes can improve the rate of the recommendation algorithm. For different data sets, when the data set is large, the increase of cluster nodes can bring significant rate to the algorithm. When the number of nodes in the cluster is relatively small, the rate of the algorithm is not significantly improved. The specific reason is that a small number of cluster nodes may be responsible for the management and scheduling of cluster data resources in addition to the data processing tasks of cluster computing. However, when the number of cluster nodes is greater than 3, the growth of acceleration ratio curve is close to linear.

4.2. MAE Values of the Algorithm for Different Neighbor Numbers

| Table 1. MAE values of the algorithm for different neighbor numbers |
|---------------------------------------------------------------|
|                   | 5    | 10   | 20   | 30   | 40   |
| Logistic          | 0.703| 0.691| 0.698| 0.701| 0.699|
| User clustering   | 0.738| 0.713| 0.722| 0.719| 0.718|

As shown in Table 1, it can be seen that when the number of neighbors increases, the MAE value of the algorithm in this paper decreases continuously and tends to be in a stable state. The MAE value of the proposed algorithm is significantly lower than that of the other two algorithms, indicating that the proposed algorithm has a good effect in terms of recommendation accuracy.

4.3. MAE Values of the Algorithm under Different Data Sets

| Table 2. MAE values of the algorithm under different data sets |
|---------------------------------------------------------------|
|                   | 5000 | 10000| 15000| 20000| 30000|
| Logistic          | 0.82 | 0.78 | 0.73 | 0.68 | 0.84 |
| User clustering   | 0.94 | 0.92 | 0.91 | 0.88 | 0.57 |
| This method       | 0.49 | 0.41 | 0.37 | 0.38 | 0.38 |

As shown in Table 2 and Figure 2, MAE values of different algorithms are different in different data sets. In the case of the same data set, the MAE value of this algorithm is significantly smaller than that of the other two algorithms. In the case of different data sets, the MAE value of this algorithm is also smaller than that of the other two algorithms. The smaller the MAE value, the more accurate the recommendation. Therefore, the recommendation algorithm proposed in this paper based on Hadoop
big data analysis platform can have a good recommendation effect under the current big data application scale.

5. Conclusions
Aiming at the change of user interest in collaborative filtering recommendation algorithm, this paper improves the collaborative filtering recommendation algorithm, combines the similarity of user interest with the traditional similarity calculation method, and proposes a measurement method based on user interest. Will be based on user interest degree improved similarity measure method is applied to the Hadoop platform, parallel processing based on graphs model, build a multi-node distributed cluster, through the contrast experiment and analysis, validate the proposed parallel computing method is more suitable for applied in large-scale data sets, effectively improve the recommended rate and accuracy, improve the effectiveness of the algorithm.

References
[1] Hu Q Y, Zhao Z L, Wang C D , et al. An Item Orientated Recommendation Algorithm from the Multi-view Perspective [J]. Neurocomputing, 2017, 269(dec.20):261-272.
[2] Saeed M, Mansoori E G. A new slope one based recommendation algorithm using virtual predictive items [J]. Journal of Intelligent Information Systems, 2018, 50(3):527-547.
[3] Barman D, Sarkar R, Tudu A , et al. Personalized query recommendation system : A genetic algorithm approach[J]. Journal of Interdisciplinary Mathematics, 2020, 23(2):523-535.
[4] Melorose J, Perroy R, Careas S. Hadoop definitive guide [J]. O'rilly Media Inc Gravenstn Highway North, 2015, 215(11):1 - 4.
[5] Goiri I, Bianchini R, Nagarakatte S, et al. ApproxHadoop: Bringing Approximations to MapReduce Frameworks[J]. Computer architecture news, 2015, 43(1):383-397.
[6] Reyes-Ortiz J L, Oneto L, Anguita D . Big data analytics in the cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf [J]. Procedia Computer Science, 2015, 53(1):121-130.
[7] Harrison, Guy. Google, Big Data, and Hadoop [J]. 2015, 10.1007/978-1-4842-1329-2(Chapter 2):21-37.
[8] Xiaojun L. An improved clustering-based collaborative filtering recommendation algorithm [J]. Cluster Computing, 2017, 20(2):1281-1288.
[9] Xu G, Tang Z , Ma C , et al. A Collaborative Filtering Recommendation Algorithm Based on User Confidence and Time Context [J]. Journal of Electrical and Computer Engineering, 2019, 2019(6):1-12.
[10] Ye S. Application of collaborative filtering algorithm in the design of library recommendation system in university library [J]. Revista de la Facultad de Ingenieria, 2017, 32(14):348-353.