Design and Development of E-commerce Recommendation System Based on Big Data Technology

Jing Gao(✉)

Shandong Vocational College of Science and Technology, Weifang 261000, Shandong, China
819018365@qq.com

Abstract. E-commerce recommendation system is the key to solve the problem of information explosion and information overload faced by consumers in the process of online shopping, and it is also an important means to tap the potential needs of consumers. Faced with the shortcomings of the current e-commerce platform recommendation system, such as low accuracy, single recommendation scheme and lack of in-depth analysis, this paper will focus on the recommendation algorithm, and integrate the conventional Item-CF and Use-rCF algorithms with the help of K-means clustering algorithm to improve the adaptability of collaborative filtering recommendation algorithm. In addition, a distributed data processing server will be built with the help of Hadoop framework to collect and store massive data, and the recommended algorithm will run smoothly with the help of Spark distributed computing engine. The overall deployment of the recommendation system will be between the user I/O interface and the e-commerce platform, subject to the call and control of the e-commerce platform Web Server. The test results show that the system has improved the recommendation efficiency and accuracy to some extent, and made a useful attempt to promote the intelligent development of e-commerce recommendation service.

Keywords: big data technology · electronic commerce · recommendation algorithm · Hadoop · computer application

1 Introduction

With the rapid rise and rapid development of network information technology, e-commerce platform uses the advantages of data resources to provide efficient and convenient services for consumers, but also brings information explosion and information overload to consumers. [1] In order to break through the influence of information overload of e-commerce platform on consumers' shopping decisions and further meet users' diversified consumption needs, e-commerce recommendation system came into being. E-commerce recommendation system can directly participate in the user's use process, aiming to help users quickly obtain the required goods, shorten the browsing and searching time, and successfully complete the purchase activities through statistical analysis of user's personal information, historical consumption records, purchase preferences

V. Gaikar et al. (Eds.): ECIT 2023, AHE 18, pp. 36–40, 2023.
https://doi.org/10.2991/978-94-6463-210-1_6

and other data information, thus improving user satisfaction. [2] The key to realize the function of recommendation system lies in the design of recommendation algorithm. At present, the common recommendation algorithms include rule algorithm, content algorithm, collaborative filtering algorithm and hybrid algorithm. [3] The four algorithms all have some limitations in practical application, which leads to the problems of low accuracy of recommendation system, single recommendation scheme and lack of in-depth analysis. In view of this, this paper believes that under the environment of big data technology, building an e-commerce comprehensive data analysis and processing server with Hadoop framework as the core and Spark distributed computing engine is conducive to strengthening the analysis and operation ability of recommendation system for massive data, and greatly improving the operation efficiency of recommendation algorithm. In addition, the system will optimize the recommendation algorithm with K-means clustering mining technology to further enhance the recommendation effect of e-commerce recommendation system.

2 Development Process

First of all, the Hadoop framework adopts cluster deployment mode, which needs to be considered from two aspects: hardware equipment and software programs. In terms of hardware equipment, according to the functional requirements of the recommendation system and the data volume of the e-commerce platform, Hadoop cluster includes three nodes, named as Master1, Slave1 and Slave2 respectively. Among them, Master stands for master node and Slave stands for slave node. Each node needs a 4-core hexadecimal CPU with 16G memory and 512G hard disk to meet the distributed storage requirements of various types of data. [4] As for the software program, Linux is selected as the bottom operating system of each node, the version is CentOS 6.8 (x86_64), jdk-1.8.0_201-linux-x64 is selected as the JDK version, and Hadoop framework version is 2.7.2. After the deployment of Hadoop cluster is completed, the Spark distributed computing framework is installed and deployed on each Hadoop node. Spark version is 2.1.1, and the parallel computing engine of big data is built together with business database MongoDB, data caching tool Redis, Zookeeper cluster resource management framework, Flume-ng log capture tool and Kafka message sequence system [5].

Secondly, all kinds of recommendation algorithms are implemented under the Spark framework. Spark's greatest advantage comes from its own ResilientDistributed Datasets (RDD). RDD is called only when using the action operator, which can greatly reduce computing resources and perform well in data mining or iterative calculation. [6] During the implementation of various recommendation algorithms, it mainly involves three parts: data reading, recommendation algorithm calculation and product similarity calculation. User information data, commodity information data and user behavior data in MongoDB will be read, and the data will be preprocessed by RDD's map operator. Input the preprocessed data into the recommendation algorithm to complete the calculation and output the commodity recommendation list TOP-X, and then record it as array [] through the map operator of RDD, and return it to MongoDB to complete the data persistence. [7].

Finally, you need to package the Spark program, submit it to YARN Explorer, and complete the deployment in Client mode. The following is the key code of the Spark task

submission command. Through the introduction of the above key technical theories, the overall environment of system development, the configuration of related software and tools are determined, and the technical feasibility of the overall project of e-commerce recommendation system is also clarified.

./bin/spark- submit.
- class org. Apache. Spark. Examples. MainTest.
- master yarn.
- deploy-mode client.
- executor-memory 512 m.
- total-executor-cores 1.
~ jars/spark- examples. Jar.

3 Functional Implementation

3.1 Data Collection Module

When the user logs on to the e-commerce platform, the personalized recommendation system will automatically start, and obtain user information, user historical consumption records and current interactive operation behavior. User information and user historical consumption records are structural data, which can be directly called by Spark SQL to MongoDB. However, the interactive operation behavior needs Flume-ng tool to obtain the user operation log of electronic system, and Kafka component is used to preprocess the log file information to obtain the necessary data for recommending algorithm.

3.2 Recommended Engine Module

Under this function module, the collaborative filtering recommendation algorithm is optimized and reconstructed. Collaborative filtering algorithm includes user-based collaborative filtering (User-CF) and item-based collaborative filtering (Item-CF). The basic operation process consists of five steps: expressing user preference information, calculating similarity, finding nearest neighbor set, predicting products and completing recommendation. [8] In the actual application process, the problems of sparse original data, poor expansibility and insufficient accuracy directly affect the functional realization of e-commerce recommendation system. Therefore, this system will use K-means clustering algorithm to improve collaborative filtering recommendation algorithm, and form a dual collaborative filtering algorithm based on users and items.

With the addition of K-means clustering algorithm, users and commodities can be clustered separately in the similarity calculation link, and the search range of subsequent nearest neighbor sets can be narrowed. Then the two calculation results are weighted and averaged, and finally the product recommendation is completed according to TOP-X. The criterion function of K-means is shown in Formula 1, where E is the sum of squared errors of all objects, p is the data point of a certain cluster, and m_i represents the center of the cluster c_i. [9] After data simulation, the performance of User-CF, Item-CF and the dual collaborative filtering algorithm based on users and commodities proposed in this paper are compared in the same neighbor set environment, and the accuracy of prediction

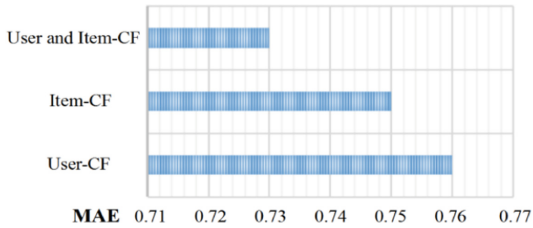

Fig. 1. Recommended quality comparison test results of the three algorithms

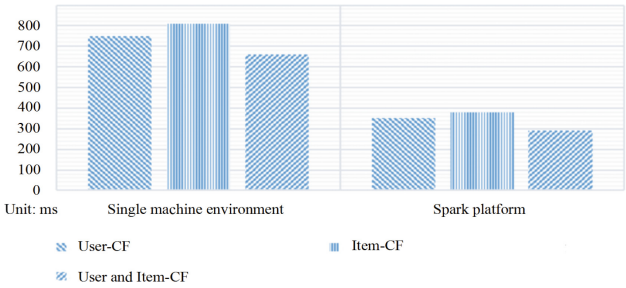

Fig. 2. Running time test results of three algorithms

is measured by the mean absolute error (MAE). The experimental results are shown in Fig. 1. The calculation of MAE index is shown in Formula 2, where p represents the real value of users' preference for goods, a represents the predicted value of system recommendation, and N represents the sum of p values. The smaller the MAE value, the higher the accuracy of system recommendation [10].

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{1}$$

$$MAE = \frac{\sum_{1}^{N} |p_i - a_i|}{N} \tag{2}$$

In addition, in order to verify the actual performance of the Spark computing engine under Hadoop framework, the running time of the three algorithms in the stand-alone environment is compared with that in the Spark platform. The test is based on the same number of concurrent operations, from the user login to the final recommendation, and based on the time stamp information of data persistence in MongoDB database, the running time of the three algorithms is counted. The test results are shown in Fig. 2.

4 Conclusion

In order to improve the performance of e-commerce platform recommendation system, on the one hand, this paper uses K-means clustering algorithm to upgrade and optimize the recommendation algorithm, improve the recommendation accuracy and solve the

expansibility problem. On the other hand, Hadoop framework and Sprak computing engine are used to improve the efficiency of recommendation system. The test shows that the system has improved the recommendation efficiency and accuracy to some extent, and made a useful attempt to promote the intelligent development of e-commerce recommendation service.

References

1. Qi Lili, Zhao Rui. Influence of Information Overload on Online Consumers' Shopping Decision[J].Journal of Commercial Economics. 2018.05.
2. Zhou Yanrong. E-commerce Intelligent Recommendation System Based on Personalized Characteristics[J].Modern Electronics Technique. 2020.09.
3. Qin Chong, Zhao Tiezhu, et al. Overview of Research and Development of Personalized Recommendation Algorithms[J].Journal of Dongguan University of Technology. 2021.06.
4. Shi Fangxia, Gao Yi. Application Analysis of Hadoop Big Data Technology[J].Modern Electronics Technique. 2021.09.
5. Jiang Yongdu, Cheng Desheng, et al. Big Data Computing Platform Based on Spark Framework[J].Network Security Technology & Application. 2020.03.
6. Cao Shijiu. Research and Implementation of Data Hybrid Computing Platform Based on Spark[D].Beijing University of Posts and Telecommunications. 2019.06.
7. Cheng Jie. Research on Commodity Recommendation Algorithm Based on Spark Big Data Platform[D].Heilongjiang University. 2022.05.
8. Liu Hualing, Guo Yuan, et al. Research Progress of Similarity Algorithm in Collaborative Filtering[J].Computer Engineering and Applications. 2022.04.
9. Wu Tingting, Li Xiaozhong, et al. Improved Collaborative Filtering Algorithm Based on K-means[J].Journal of Tianjin University of Science & Technology. 2021.12.
10. Qin Qionghua. Research on Personalized Recommendation System Based on Collaborative Filtering Algorithm[J].Science & Technology Information. 2022.05.