Collaborative Filtering Algorithm on the Basis of Hadoop Corn Seed Optimization

Haiping Bi [a*] and Dongming Li [b]

School of Information Technology Jilin Agricultural University Changchun

[a]603944597@qq.com; [b]23695699@qq.com

Keywords: Corn; Seed; Optimization; Hadoop; Collaborative; Filtering; Algorithm

Abstract. This Seeds are the most basic and primary production materials in agricultural production. How to select suitable seeds for planting is the key to increase agricultural yield and harvest. However, in recent years, farmers have a difficult problem in corn seed selection process. Due to the complicated factors affecting the planting of corn seeds, the traditional corn seeds selection platform will rarely combine the geographical information and the common characteristics of farmers, it can’t make effectively recommendations for corn seeds that is really suitable for farmers to plant. At the same time, the traditional data processing mode has the problem of poor throughput and low load capacity while performing operation and storage. By summarizing the above problems, the paper designs the optimized management platform of corn seed based on private cloud by studying the method of hadoop distributed storage and collaborative filtering algorithm. The platform is based on big data-related technology to collect, store and analyze data, to solve the problems of low resource efficiency and insufficient storage capacity. Experimental data shows that the mean absolute error(MAE) of the recommended results is reduced by 14% compared with the traditional algorithm, and the recommendation quality is higher than the traditional algorithm.

Introduction

In recent years, with the research and development of breeding technology by leaps and bounds, the number of seed-related businesses is growing, and varieties update too fast, the varieties of their respective operations are multifarious, it makes a growing number of varieties of seeds. At the same time, due to the factors which influence the production of corn are complex and changeable, farmers are lack of professional knowledge understanding on the land information and the seed information, they can't provide accurate data, it led difficult problems to the farmers when encounter choosing seeds selection[1]. In terms of enterprises, due to the complexity of the crowd to buy seeds, and the territory is numerous, seed suppliers can’t get the planting situation related information of farmer’s feedback timely and accurately, so it influences the development and production of seeds. Traditional single-node system architecture to compute and storage corn seeds data, there will be defects such as the strong dependence, poor scalability in storage and computing, weak system throughput, which greatly reduce the overall ability to deal with the data of the system. Hadoop is a distributed parallel computing model with high throughput and high fault tolerance[2]. Therefore, a distributed management solution based on Hadoop becomes a feasible means to solve this problem[3].The recommendation algorithm[4] can be recommended for the target user's preferred project according to the information entered by the user, and it can provide help for users to find the information what they need. Currently, there are two recommendation algorithms commonly used in personalized recommendation techniques: content-based recommendation[5] and collaborative filtering recommendation. The principle of content-based recommendation is through the identification of the user's unchanged interest, after constructing the user preferences project characteristics in according to the user's historical data, then fitting the user information features and content features to recommend projects to target users. Collaborative filtering algorithms are mainly divided into two algorithms based on user and project[6-7].The basic principle is: by calculating the similarity between each base user and target user's project score, finding the nearest neighbor to the target user, and making recommendation to the target user based on the recent

DOI: 10.25236/icess.2019.172

neighbor rating[8-9]. But there is a problem of the traditional recommendation algorithm in corn seed optimization recommendation, such as it can only according to the farmers' seed selection historical record information to recommend the similar projects that match preferences, while ignoring the factors related to geographical characteristics. Through the research on traditional collaborative filtering algorithm, it advances the collaborative filtering algorithm based on geographical features, by increasing the similarity between farmers to increase the accuracy of recommended seeds for the target farmers, to make the optimal recommendation to selection of seeds for farmers in the optimization process. In this article, it develops the seed optimized private cloud platform based on the Hadoop parallel computation, and joins the collaborative filtering algorithm based on geographical environment characteristics. It provides a kind of thought and method for solving the complex agricultural data and provides a reliable theoretical basis for constructing intelligent agriculture.

Key Technologies for the Platform Implementation

Hadoop Computing Model. Hadoop is an open source large data distributed storage platform owned by the Apache software foundation. Compared with traditional single-node processing mode, as a software framework that can do the distributed processing of large amounts of data, Hadoop mainly has the following advantages, High reliability. In the Hadoop distributed storage platform, when the data is sent to a single node, it also be copied to other nodes of the cluster, and can automatically redistribute failed tasks. Efficiency. Hadoop can move data dynamically from one node to another, while ensuring the dynamic balance of each node, so processing data is very efficient. High scalability. Hadoop distributes data among idle computer clusters and finishes computing tasks, which can be extended to thousands of nodes. This also determines that hadoop can be extended to more cluster nodes. High fault tolerance. Hadoop stores data automatically in as many copies as it can, and when a failure occurs, it can redistribute the failed task.

MapReduce Parallel Computing Model. MapReduce is a distributed computing framework model for solving massive data[10]. The MapReduce computing framework will store the unstructured data calculated in Hbase or HDFS, and structured data which like user information can be stored in relational databases. MapReduce abstracts the complex, parallel computing processes that run on large clusters to two functions: Map functions and Reduce functions. MapReduce splits a large data set stored in a distributed file system into separate slices that can be processed in parallel by multiple Map tasks. The process as shown in figure 1.

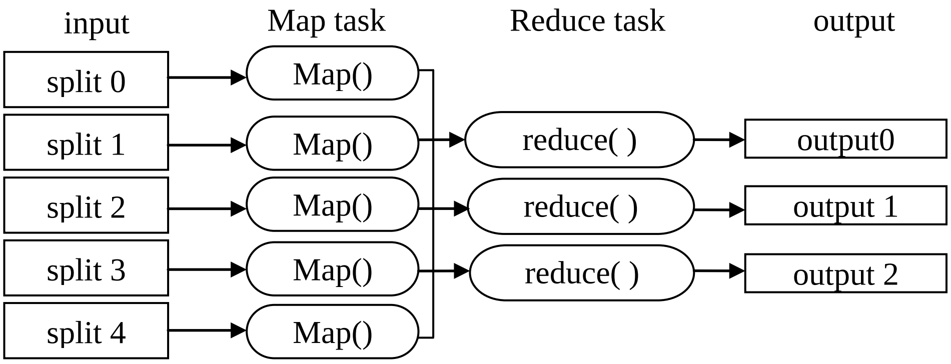

Figure 1. MapReduce working flowchart

Traditional Collaborative Filtering Algorithm. Collaborative filtering algorithm is a widely used personalized recommendation technology, through searching the most adjacent user that is similar to the target user who is graded on the event to produce prediction recommended to him, how to accurately find the target user nearest neighbors is the key to the problem, how to find the nearest neighbor of the target user is the key to the problem.

Collaborative filtering algorithms mainly include three types, Collaborative filtering algorithm based on the model, the study of historical data forms a model, and the model is used for prediction[11].Collaborative filtering algorithm based on in-memory, by calculating a similar statistical method to generate the nearest neighbor set that has similar interests to the target user. Collaborative filtering algorithm based on the project, the statistical method is used to calculate the similarity of the predicted and graded item, and the graded item is weighted to get its prediction score[12].The collaborative filtering algorithm can be divided into three steps from the principle of work: data expression, query of similar user sets, and generation of recommended data sets. Data expression. Supposed that in a recommendation system there are m users and n projects, the system can be expressed as a matrix m*n, then each item represents user 's rating of item I in the current matrix. Query the similar user sets. In this step, by calculating the similarity between the target users and the users in the matrix to be graded, looking for the similar sets as similar neighbors.Through the modified cosine similarity computing similarity between users: setting to be collection of items that have been graded together by users I and j ,the set of items that user I and user j have been graded are represented by and respectively, the similarity sim (I, j) between user I and user j is:

$$\mathrm{sim}(i,j)=\frac{\sum_{c\in I_{ij}}(M_{i,c}-\overline{M}_{i})(M_{j,c}-\overline{M}_{j})}{\sqrt{\sum_{c\in I_{i}}(M_{i,c}-\overline{M}_{i})^{2}}\bullet\sqrt{\sum_{c\in I_{j}}(M_{j,c}-\overline{M}_{j})^{2}}} \tag{1}$$

Generating recommended data sets. Through similar neighbors focusing on users interest to make recommendations to object user.

$$P_{\mathrm{u},\ \mathrm{i}}=\overline{M_{\mathrm{u}}}+\frac{\sum_{a=1}^{\mathrm{n}}(M_{a,i}-\overline{M_{a}})*sim(u,a)}{\sum_{a=1}^{n}\left|\mathrm{sim(u,a)}\right|} \tag{2}$$

$\overline{M_{\mathrm{a}}}$ and $\overline{M_{\mathrm{u}}}$ are represented as the average score of item by user u and user a respectively,sim（u, a）is the similarity coefficient, $M_{\mathrm{a},i}$ is represented as the score by user u to item i, n is the number of users.

Platform Structure and Function

Big Data Processing Strategy of Seed Optimization. The first task of seed optimization platform is to realize the unified organization and management of massive seed data. It mainly refers to data acquisition, data processing, data information processing, data information parallel computation, data analysis, distributed storage, and so on. The process as shown in figure 2.

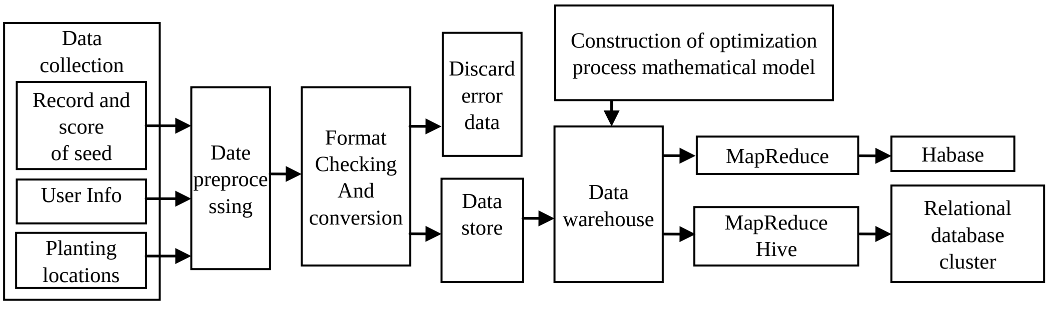

Figure 2. Seed optimization processing strategy

In the data collection phase, through the information collection, such as user information, seed usage records and scoring, the main farming methods and plant location, the structured data can be obtained, and data capture-dis are stored in HDFS to build the data warehouse after the data classification, cleaning, removing duplicate. Through MapReduce and Hive methods to do data analysis to the source data, constructing the seed optimization data model, recommending the optimal results to farmers and giving information records to the seed suppliers.

Basic Data Collection. There are many factors that influence the yield of corn, selecting appropriate information can improve the accuracy of corn seeds recommendation. On the basis of data collection, selecting farmers' information, crop planting place, seed use records and scoring information which have the characteristic of relatively objective stability and easy collection to collect, which can reduce the difficulty of using seed selection platform. User information. Through collecting and recording the farmer's name and contact information, it is convenient for breeding companies to establish communication channels with farmers to solve technical problems in the process of cultivation and provide timely feedback on product information. Use records and scoring of seeds. Making the survey on farmers in nearly three years at the use of corn seeds, the seeds' name, yield and scoring information are recorded, and the score is divided into {1, 2, 3, 4, 5} five grades, which can be graded according to the scoring. The location of the plant. The location information of the farmer's crop is recorded and divided according to the provinces, prefectures, counties, townships and villages. The main farming methods. The main agricultural machine information used in farming is recorded in order to assess farmers' ability to develop the land.

Optimization Process Data Model. In order to find similar characteristics of farmers set, through seeds use records and farmers' scoring to find similar users' set, the introduction of planting location information can be used to eliminate the interference of unrelated farmers from the target farmers, so as to find more accurate collection of nearest neighbor users set[13-14]. Establishing the information data table of farmers, which records the farmer's name, contact information, crop planting record and grading, crop planting location, main farming methods and other information. Calculating the farmers' similarity sim1 based on the farmer's characteristic data, and the calculation procedure is shown as follows. The farmer attribute matrix of m * n is U = [m, n], is the j feature attribute of farmer a, is the j feature attribute of farmer b, if the j feature attribute of farmer a and b have the same value is 1, then the feature correlation between farmer a and b is to calculate the ratio of same features of a and b in the total number of farmers, the formula is calculated as follows:

$$sim1(a,b) = \sum Similar(U_{aj}, U_{bj}) / n \tag{3}$$

Calculating the farmers' final similarity, In the modified cosine similarity formula, the farmers' geographical feature parameters are introduced, the improved farmers' similarity sim (I, j) based on the farmers' geographical characteristics is obtained, the calculation equation is shown as follows:

$$\mathrm{sim}(i,j) = \frac{\sum_{c\in I_{ij}} (R_{i,c} - \overline{R_i})(R_{j,c} - \overline{R_j}) sim1(i,j)}{\sqrt{\sum_{c\in I_i} \left(R_{i,c} - \overline{R_i}\right)^2 sim(i,j)} \bullet \sqrt{\sum_{c\in I_j} (R_{j,c} - \overline{R_j})^2 sim(i,j)}} \tag{4}$$

After joining the geographical features to find the farmers' set with similar characteristics, the similar farmers are brought into the recommendation algorithm to predict the similarity among the more similar farmers[15-16].

$$\mathrm{p}_{u,i} = \overline{M_u} + \frac{\sum_{a=1}^{n} (M_{a,i} - \overline{M_a}) * sim(u,a)}{\sum_{a=1}^{n} \left|sim(u,a)\right|} \tag{5}$$

Experimental Results

In order to verify the collaborative filtering algorithm based on geographical position characteristic performance, data from seed sites are used in the experiment, which include 60000 scoring records from 956 farmers to 2643 corn seeds, the farmers' information, the seed use records and scoring, the main farming methods and planting sites and so on. In the course of the study, data sets are divided into training sets and test sets by 70% and 30%.

The mean absolute error (MAE) method is adopted to evaluate the deviation between the predicted value and the actual value. The smaller the deviation, the higher the prediction accuracy is, the higher the quality index recommended by farmers is. The farmers' sets that are predicted is represented as { $p_1, p_2, ..._1, p_n$ }, and the corresponding sets of actual farmers' scores is { $q_1, q_2, ..._1, q_n$ }, the average error MAE is defined as:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \qquad (6)$$

In the test, the number of the most neighbors is ranged from 5 to 40, the interval is 5, and the experimental results are shown in table 1. Improved collaborative filtering algorithm(GCF) with the traditional algorithm based on the ICF algorithm and the MAE of UCF based on the user are shown in table 1.

Table 1 Styles improved collaborative filtering algorithms and traditional algorithms MAE

neighbors	**MAE**		
	UCF	***ICF***	***GCF***
5	0.8714	0.8153	0.7842
10	0.8542	0.8061	0.7918
15	0.8445	0.7734	0.7627
20	0.8338	0.7612	0.7416
25	0.8222	0.7516	0.7296
30	0.8116	0.7449	0.7139
35	0.8043	0.7517	0.6865
40	0.7951	0.7411	0.6699

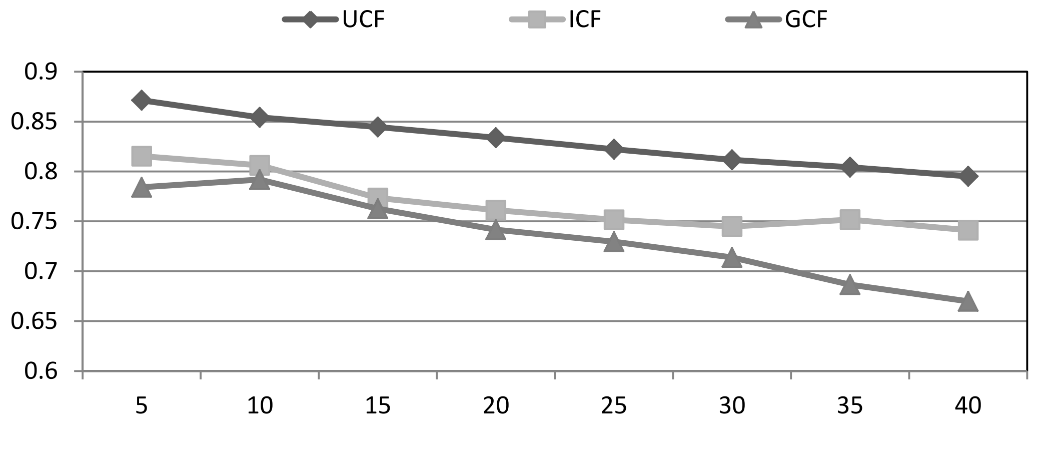

Figure 3. The Comparison of MAE of the proposed algorithm

By the information in table 1 and figure 3 can be summarized while the average absolute error of the recommended results of collaborative filtering algorithm based on geographic information is compared with the traditional collaborative filtering algorithm, MAE compared with the traditional collaborative filtering recommendation algorithm UCF is reduced by 14%,and reduced by 9% than ICF , because the smaller value of the MAE algorithm is, the higher recommended precision is, it can be concluded that compared with the traditional collaborative filtering recommendation algorithm, the recommendation accuracy of collaborative filtering algorithm based on geographical features is higher.

The main reason is when the collaborative filtering algorithm based on geographical features search for similar farmers' set, it joins the crops planting and other relevant information, which makes higher accuracy when looking for similar farmers. By increasing the number of nearest neighbors, the new information of crop is increasing, and the value of MAE is constantly decreasing, it is proved the accuracy and effectiveness of the collaborative filtering algorithm based on geographical features.

Conclusions

This paper designs and realizes the corn seed optimization data processing model based on Hadoop, through the seed data acquisition, parallel computing and distributed storage process, the Hadoop distributed computing framework and collaborative filtering algorithm is applied to corn seed optimization platform, and joining the geographic information characteristics above the traditional collaborative filtering algorithm, obtaining more precise nearest neighbors set, and solving the problem of the selection of corn, the enterprise can also collect the feedback information of farmers more timely and comprehensively. The experimental results show that the proposed algorithm can improve the performance of recommendation effectively. This paper provides a thought and method for dealing with the complex agricultural data and provides a reliable theoretical basis for the construction of intelligent agriculture.

References

[1] L.T. Weng, Y. Xu and Y.F Li. An Improvement to Collaborative Filtering for Recommended Systems.[C] //Proceedings of the 2005 International Conference on Intelligent Agents, Web Technological and Internet Commerce . Washington: IEEE Computer Society, 2005:792-795.

[2] J.Liu Design and research of massive teaching resource storage platform based on Hadoop[J]. Computers and Telecommunications, 2013, (07): 27-29.

[3] F. Yang, H.R. Wu, H.J.Zhu, H.H. Zhang and X. Sun. Massive agricultural data resource management platform based on Hadoop[J]. Computer Engineering, 2011, (12): 242-244.

[4] Sarwar B, Kary pis G, Ko nstan J. Item-based Collaborative Filtering Recommendation Algorithms. [C] //Proceedings of the 10th International WWW Conference . New York: ACM, 2001: 285-295.

[5] L. Wang, and Z.J. Zhai. A collaborative filtering algorithm based on time weighted[J]. Computer application, 2007(09): 2302-2303.

[6] T. Li, and J.D. Wang. A collaborative filtering recommendation algorithm based on user clustering[J]. Journal of software, 2007(07): 1178-1183.

[7] H.Z. Yang, X.Q. Cong, and M.L.Liu. Research on Personalized Recommendation Algorithm Based on time weighted[J]. Computer engineering and Science . 2009, 131(16): 126-128.

[8] D.W. Peng, and B. Hu. A collaborative filtering algorithm based on user characteristics and time[J]. Journal of Wuhan University of Technology, 2009, 31(3): 24-28.

[9] C. Li, Z.M. Zhu, X.F. Gao, and Y.F. Chen. Neighborhood decision based collaborative filtering recommendation algorithm[J]. Computer Engineering, 2010(13): 135-138.

[10] W.F. Zhang. The principle and design of distributed computing platform based on Model [D]. Wuhan: HuaZhong University of Science and Technology, 2010.

[11]L.Y. Qiu, W.D. Qiu, Q. Su and L. Liao. Implementation of distributed hash algorithm based on Hadoop[J]. Information security and communication security. 2011, (11): 54-56.

[12]X.S. Ji, Y.B.Liu, and L.M. Luo . Similarity measurement method based on user's interested in collaborative filtering [J]. Computer application, 2010(10): 233-237.

[13] Q. Wang, and J.B.Wang. An improved collaborative filtering recommendation algorithm[J]. computer science, 2010(06): 53-57.

[14]F.G. Zhang. Research on trust based collaborative filtering algorithm with multiple user interests[J]. Mini-Micro Systems, 2008(08): 112-116.

[15] M. Xiao, and Q.X. Xiong. Collaborative filtering recommendation algorithm based on semantic similarity of project[J]. Journal of Wuhan University of Technology, 2009, 31(3): 21-23.

[16] M.L. Liu,T.C. Liu, and F Zhang. Recommendation algorithm for project scoring prediction based on bidirectional association rules [J]. Journal of Wuhan University of Technology, 2011, 33(9): 150-155.