Intelligent Courses Recommendation System of Collaborative Filtering Algorithm Based on K-means Clustering under Spark Platform

Zhaowei Chen

Hangzhou Dianzi University
**Corresponding author. Email: 605180451@qq.com*

ABSTRACT

Students in Chinese universities face a variety of electives. Their choice of the course impacts their scope of knowledge and their grade point, which influences their career or further learning. But in most condition, the most appropriate option will not be chosen due to the lack of detail and reference.

Intelligent recommendation system was applied in multiple fields, collaborative filtering algorithm is one of common recommendation algorithm and is widely used. However, a collaborative filtering algorithm has the drawback in accuracy and user similarity. In that case, an improved algorithm based on K-means clustering is applied. The improved algorithm uses Dichotomous K-means algorithm to deal with the long distance of cluster centroid. As original similarity algorithm leads to the result of low user similarity, a mixed algorithm is applied.

Traditional methodology of data computing and data storing cannot handle the demand of that much data set. As a big data platform, Spark has become a popular solution to these problems. University management combined with big data is a tendency[4].

Keywords： *Spark，Elective，Collaborative filtering recommendation, K-means.*

1. INTRODUCTION

Intelligent recommendation system was applied in multiple fields,such as movie recommendation [1],items recommendation in internet [2].The introduction of recommendataion system could help a lot in students to select elective. The application of K-means algorithm improve the accuracy of the system in the base of collaborative filtering algorithm [3] hence accomplish better performance in recommendation. Chinese university have huge scale of students records with an increasing trend of freshmen enrollment.

Inuniversity course system of China, students can only select the elective without an intelligent recommendation. The first impressions tend to be the only reason for selecting, not a recommendation of individuation from the system.

Spark platform has a good performance in real-time reaction. The characteristic that its process in the memory makes it more advanced in efficiency. It will be the key of a recommendation system which need high real-time performance.

2. ANALYSIS THE ELECTIVE SYSTEM BASED ON SPARK

2.1. Spark

As a computing engine, the spark can integrate with many applications, data sources, store platforms and environment using data processing ability in memory and convenient API. In spark, advanced API of Java,Python,R is available.It can be used in big data processing task, such as streaming applications, machine learning and also interactive SQL searching in huge data set which need to be requested iteratively.

The prime advantages of Spark are its high efficiency of processing, multiple API and unified big data engine with handy components. Compared with Map-Reduce model, Spark has real-time structure which is easier to program and faster.

2.2. Spark RDD

Spark RDD (Spark Resilient Distributed Dataset) is the abstract representation of data.

RDD is the record set of a read only partition. Creating operation of a RDD can be done in a stable storage or other RDD data. RDD is also a fault-tolerate element set which can be parallelly operated.

RDD is immutable, every data set can be divided so that be dealt with in different node. RDD contains every type of Python, Java, Scala object and the class defined by user.

3. ELECTIVE IN CHINESE UNIVERSITY

Chinese undergraduates have two options when they graduate: applying for jobs or further learning. Applying for jobs need more practical experience and knowledge, enterprises prefer applicants who have already mastered professional skill so that they can finish their job perfectly, so the students who want to apply for jobs after graduating tend to select the elective which related to their favored jobs.

Professional elective focus on a certain topic, the knowledge of professional elective can only be used to deal the problem in a segmentation fields. Take Computer Science as an example, network security is an elective of Computer Science, but not all programmers work in network security but in developing web. Students will be benefited for choosing the elective which offer the professional knowledge required by their future jobs. However, the difference between occupations in the same field is hard to distinguished, leading to that students having no idea of their choices. In this case, students tend to look for advise from upperclassmen who have graduated in the same major of same university. Recommendation system can take the responsibility of upperclassmen to give advise if students do not know which occupation is the most suitable for them, after they decide on the job orientation, it is easy for students to select elective.

In Chinese university, students require enough credit forprofessional elective so that they can graduate successfully. Chinese undergraduates have two options when they graduate: applying for jobs or further learning. On one hand, applying for jobs need more practical experience and knowledge, enterprises prefer applicants who have the professional skill so that they can accomplish their job earlier, the students who want to apply for jobs after graduating tend to select the elective which related about their favored jobs. On the other hand, while applicating for further learning, university requests a higher grade of the study, especially the score of professional elective. Therefore, these two different options have different demands, author analysis these two options together so that apply for a weighting algorithm, making students can set the weight of job hunting and further learning.

The choice of elective is such significant, the right decision can benefit student in study or career. While, in fact, many students cannot make the right decision due to the lack of social experience or academic knowledge. For the students who prefer applying for a job directly, they still have no idea of where their interest is.

All the students in the recommendation system have accomplished most of their required course, the record can be the input of the algorithm to portray the user picture of students.

The record includes the interest grade given by students representing how much interest they have in this course and academic grade given by teachers representing how much knowledge did the student grasp. By analysing these records of user students and all the upperclassmen. The algorithm will divide the user student into the group of upperclassmen who have the highest similarity with he or her. And then recommend him or she the course get the highest weighted arithmetic mean (weight given by user student).

However, in terms of grade, everyone has a different ability of learning, someone may get a higher grade in all courses, so the input of the algorithm will be the difference value between the grade and the average grade of every student, the performance in courses only be defined by comparing with themselves. Both the interest, every student have a different standard to give interest grade, the final result will depend on the difference value between the grade and average grade to remove the error.

4. RECOMMENDATION ALGORITHM

Collaborative filtering has two kind of algorithm:user-based collaborative filtering and item-based collaborative filtering. The former recommend items to one user based on others who have similar preference.The latter recommend items to user based on other items which user used to like.

User-based collaborative filtering:If user1 and user2 have similar interest in some items(item1,item2),algorithm regard them as similar users and recommend one’s favour item(item3) to the other.

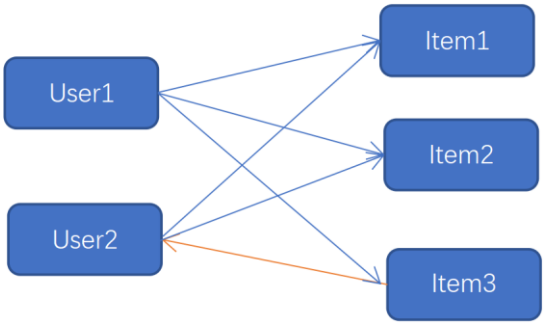

Figure 1 User-based collaborative filtering

Item-based collaborative filtering: If item1, item2, item3 have high similarity and user like item1 and item2, then the item3 will be recommended to the user.

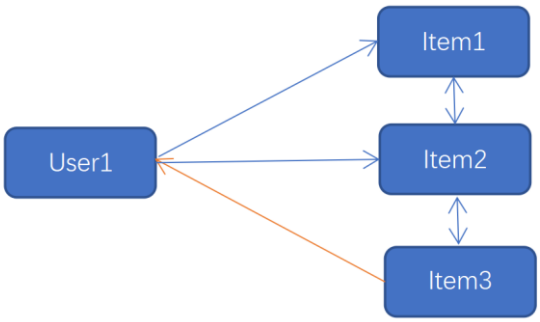

Figure 2 Item-based collaborative filtering

Intelligent courses recommendation system uses user-based collaborative filtering algorithm[6]. The system regard user who use the system to select elective as "students", who have graduate from the university as "upperclassmen",both of them are users.The process of user-based collaborative filter tend to be divided into three steps:

Forming user-item grade matrix based on historical data[7]

Users' behavioral data information is collected before. It contains the grade users get and their upperclassmen get in every required class, the higher grade users get in one class then other class, it reflect that users behave better in this course. System will compute every user's average grade point as standard value because every have different learning ability, one's worst grade may even higher than another's best.In the same way, the grades given by users depend on their interest form another matrix, higher grade represent higher interest in this matrix. The matrix store the grade of required course j, given by user i. If i have not get grade of required course j, the value will be 0.

$$R_{m*n}=\begin{bmatrix} r_{11} & \cdots & r_{1j} & \cdots & r_{1n} \\ \vdots & & \vdots & & \vdots \\ r_{i1} & \cdots & r_{ij} & \cdots & r_{in} \\ \vdots & & \vdots & & \vdots \\ r_{m1} & \cdots & r_{mj} & \cdots & r_{mn} \end{bmatrix} \quad (1)$$

Searching for closest neighbour of user by similarity computation

The recommendation list comes from favor projects of students' neighbour. Neighbour are upperclassmen who have the highest similarity with the students. The similarity is calculated based on the R_{m*n} matrix. For students, find more similar upperclassmen would get more accurate recommendation. So the way to find neighbour determine the result which depend on similarity algorithm. Universally, similarity is calculated by cosine similarity formula:

$$\mathrm{sim}(u,v)=\cos(\vec{u},\vec{v})=\frac{\vec{u}*\vec{v}}{\left\|\vec{u}\right\|^2*\left\|\vec{v}\right\|^2} \quad (2)$$

$\left\|\vec{v}\right\|$ means the norm of user v's grade vector $\left\|\vec{u}\right\|$ means the norm of user u's grade vector.

The cosine similarity calculate the cosine value of included angle of two row vectors. The smaller the angle is, the two vectors are more similar.

In the fact, when the formula comes to compute the interest grade but not the course grade. The data is more subjective, every user give their grade depend on their own standard.The similar condition is in course grade due to the difference learning ability mentioned before. As a result, the average grade point is used to adjust the formula:

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I_{uv}}(r_{ui}-\bar{r}_l)(r_{uj}-\bar{r}_l)}{\sqrt{\sum_{i\in I_{uv}}(r_{ui}-\bar{r}_l)^2}\sqrt{\sum_{i\in I_{uv}}(r_{uj}-\bar{r}_l)^2}} \quad (3)$$

In formula(3.3), I_{uv} is the intersection of the grades of user u and user v; r_{ui} is the grade given by user u of required course i; r_{uj} is the grade given by user u of required course j; $\bar{r}_l$ is the average grade of course i.

Pearson product-moment correlation is another formula to calculate the similarity:

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I_{uv}}(r_{ui}-\bar{r}_u)(r_{\mathrm{vi}}-\bar{r}_v)}{\sqrt{\sum_{i\in I_{uv}}(r_{ui}-\bar{r}_u)^2}\sqrt{\sum_{i\in I_{uv}}(r_{vi}-\bar{r}_v)^2}} \quad (4)$$

In formula(3.4), I_{uv} is the intersection of the grades of user u and user v; r_{ui} is the grade given by user u of required course i; r_{vi} is the grade given by user v of required course i; $\bar{r}_u$ is the average grade given by user u; $\bar{r}_v$ is the average grade given by user v.

(3)Getting the final recommendation result

$$P_{u,i} = \bar{r}_u + \frac{\sum_{n \in Nu} sim(u,n) * (r_{ni} - \bar{r}_n)}{\sum_{n \in Nu} |sim(u,n)|} \quad (5)$$

In formula(5): $\bar{r}_u$ is the average grade given by user u; $\bar{r}_n$ is the average grade given by neighbour n; r_{ni} is the grade of item i given by user n; $sim(u,n)$ is the similarity between user n and neighbour n.

5. K-MEANS CLUSTER ALGORITHM

As one of cluster algorithm,K-means algorithm has the advantages in easy realizing and high accuracy.K-means algorithm set a k value before clustering which means the number of cluster, then segment all the data[8].

Through the computing of similarity,the algorithm segment data into their closest original centroid of cluster which selected randomly.Then repeating this step until every centoid does not change in a precision range.

Set the data set of cluster:

$$X = \{x_i | x_i \in R^P, i = 1,2,\ldots,n\}$$

K centroid of cluster are $M_1, M_2, \ldots, M_k$ respectively.Using $W_j (j = 1,2,\ldots,k)$ as k kinds of cluster.Using Sum of Squared Error as error criterion.

Formula1:Euclidean distance between two data object:

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T (x_i - x_j)} \quad (6)$$

Formula2:arithmetic mean of same kind of data object:

$$M_j = \frac{1}{N_j} \sum_{x \in W_j} X \quad (7)$$

Formula3:cluster function:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_j} d(X_j, Z_i) \quad (8)$$

Step1:Select k sample data as original centroid of cluster randomly in X which contain N sample: $M_i (i = 1,2,\ldots,k)$

Step2:Use formula(4.1) calculate the distance between M_i and every sample data p.

Step3:Find the smallest distance of every sample data p, add p into the cluster which M_i located in.

Step4:After traversal of every sample, use formula(4.2) to calculate the value of M_i as new centroid of cluster.

Step5:Repeat step(2)~step(4) until the value of E do not change.

6. COLLABORATIVE FILTERING ALGORITHM BASE ON K-MEANS CLUSTER

Every data in algorithm consist of user,item,grade.Set the set of users as $U = \{u_1, u_2, \ldots u_m\}$,the set of users generated by K-means algorithm is $U^a = \{C_1^a, C_2^a, \ldots, C_k^a\}$,k means the number of cluster, C_k^a means cluster k.

The Collaborative filtering algorithm base on K-means cluster has steps like below:

Input: User u, User-Grade Matrix R_{m*n} of user u and all upperclassmen(grade and interest respectively),number of cluster k, the weight of interest and grade.

Output: Top-N of the electives recommended[9].

Step1:Form the final User-Grade Matrix with the weighted average of grade R_{m*n} and interest R_{m*n}.

Step2:use $M_i (i = 1,2,\ldots,k)$ as original cluster,use K-means algorithm segment R_{m*n} into k kinds.

Step3:use formula(4.1) calculate the similarity of u and k clusters, add u into the cluster with the highest similarity.

Step4:use formula(3.4) to calculate the similarity between u and other users in cluster to find the nearest neighbour $N_{uj}^a (j = 1,2,\ldots,m)$

Step5:after get the nearest neighbour, use formula(3.5) to recommend user u the Top-N electives.

7. CONCLUSION

With the development of advanced education, the choice of elective becomes increasingly important. An intelligent course recommendation system giving advice to students can not only improve students' grades but also

benefit their careers. Spark is an excellent platform to handle the large amount of data and give a real-time response, which is important in a recommendation system. The user-based collaborative filtering algorithm combined with K-means algorithm has better behavior in accuracy. It can form the user-item matrix for similarity calculating and apply the result into K-means algorithm.

However, problems still exist in the system. Data sparsity have not been well solved in the system when forming a user-grade matrix due to the case that some students may not accomplish all their required course but failed the exam before. We can fill the matrix by 0, but it will affect the accuracy of the result a lot, especially for those who did not pass a lot of course. A viable solution might be using forecast algorithm to adjust the data based on other exited data, it will be helpful for increasing accuracy.

REFERENCES

[1] Salam Salloum & Dananjaya Rajamanthri.(2021).Implementation and Evaluation of Movie Recommender Systems Using Collaborative Filtering. JAIT(3). doi:10.12720/JAIT.12.3.189-196.

[2] Qian Chenyue,Jin Yanjuan & Teng Xiaobing.(2021).Research on Commodity Recommendation System Based on Deep Learning. Journal of Physics: Conference Series(4). doi:10.1088/1742-6596/1865/4/042011.

[3] Eshetu Tesfaye,Pooja,Rani Astya Eshetu Tesfaye,Pooja & Rani Astya.(2019).Intelligent Collaborative Recommender System by Crow Search Algorithm and K-Means algorithm. International Journal of Recent Technology and Engineering (IJRTE)(2).

[4] Xu Bin.(2021).Research on The Application of Big Data in University Management. International Journal of Education and Teaching Research(1).

[5] Ting Hou (2021) Application of Collaborative Filtering Recommendation Algorithm in Smart Libraries. Technology Information(33),149-151. doi:10.16661/j.cnki.1672-3791.2111-5042-0189.

[6] Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. Knowledge-based systems, 26, 225-238.

[7] Deshpande, M., & Karypis, G. (2004). Item-based top-n recommendation algorithms. ACM Transactions on Information Systems (TOIS), 22(1), 143-177.

[8] Erisoglu, M., Calis, N., & Sakallioglu, S. (2011). A new algorithm for initial cluster centers in k-means algorithm. Pattern Recognition Letters, 32(14), 1701-1705.

[9] Deshpande, M., & Karypis, G. (2004). Item-based top-n recommendation algorithms. ACM Transactions on Information Systems (TOIS), 22(1), 143-177.