Course Selection of Students Based on Collaborative Filtering

Ren Xueli
School of Computer Science and Engineer
Qujing Normal University
Qujing, China
oliveleave@126.com

Dai Yubiao
School of Computer Science and Engineer
Qujing Normal University
Qujing, China
abiaodai@163.com

Ning Deqiong
School of Computer Science and Engineer
Qujing Normal University
Qujing, China
ningdqiong@126.com

Chen YongMei
School of Computer Science and Engineer
Qujing Normal University
Qujing, China
chenymei@163.com

Abstract—The credit system is the need of higher education development. It is the basis and core of the credit system. Therefore, it is necessary to establish a reasonable course recommendation system. Collaborative filtering is a method of group intelligence, which has been successfully applied in many fields such as electronic commerce, video recommendation, etc. A model to recommend the curriculum is established to guide students to choose the right course to improve the learning effect based on a large number of students grades in the educational management system, and that combined with collaborative filtering recommendation technology to find the students who may not pass and give any help as soon as possible, so these improve the level of management of students in the school effectively and lay a solid foundation for improving the quality of teaching. The method is applied to the prediction of student grades, the results show that the absolution deviations between predicted and actual values are within 5.

Keywords—Collaborative Filtering; Credit System; Course Selection; Student Management; MAE

I. INTRODUCTION

With the continuous expansion of colleges and universities, higher education has been transformed from outstanding mode into the quality and the mass mode, the scale of the school continues to expand, the students have a large difference in the level and the starting point, therefore the implementation of the credit system in Colleges and universities meet not only the requirements of the times, but also the needs of higher education development and law of personnel training [1]. There are the three characteristics of the credit system: elective course system, flexible educational system and target management, and the elective system are the basis and core of the credit system. The implementation of the elective course system makes the students have the sovereignty in the process of learning, and that is a very important role in optimizing the knowledge structure, promoting the development of teaching, improving the teaching content and developing students' personality. Most of the students in our country grow up in the exam oriented education, lacking the ability of self-education management, having a large of lacking in the need of society and myself cognition, so which leads to this phenomenon that the students don't know how to arrange spare time, many students neither have a clear goal, nor know what to learn, how to learn, and there is no positive learning attitude [2,3]. A direct result of lacking cognitive is that they don't select these courses meeting to their requirement of knowledge structure and professional need, some students select course according to the difficulty and so on, the appearance of these situation brings negative influence on the implementation of the elective system, therefore, it is necessary to develop a set of perfect course recommendation system. A course recommendation model based on grade prediction is established to guide students to choose courses in this paper, which has a basis on the test results of the students in student management system, combined with collaborative filtering recommendation technology.

II. COLLABORATIVE FILTERING

Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating) [4-10]. The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue x than to have the opinion on x of a person chosen randomly. The similarity between the users is measured in order to find out the nearest neighbor of the target user, and typical similarity metrics are Euclidean Distance, Cosine, Modify cosine and Pearson correlation [10-13].

A. Euclidean Distance

If uses rating look as the points in Euclidean space, then the distance in the points is similarity for them. If the common item set is Ij, which includes the items rated by user i and user j, and are the rate which are rated separately by user I and j, then the similarity between user I and user j is computed used Formula 1.

$$sim(i, j) = \frac{1}{1 + \sqrt{(\sum_{c \in I_j} (R_{i,c} - R_{jc})^2)}} \quad (1)$$

Where $R_{i,c}$ is the rate of item c by user I; $R_{j,c}$ is the rate of item c by user j.

B. Cosine

If uses rate look as the vectors in n space, then the similarity between one user and the other user is defined as cosine between one vector and the other vector. If $\vec{i}$ and $\vec{j}$ are rating vectors by user i and user j, then the similarity between user I and user j is computed used Formula 2.

$$sim(i, j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\| \|\vec{j}\|} \quad (2)$$

C. Modify Cosine

As the different user's rating scale does not consider in the cosine similarity, the modified cosine similarity is used to improve the defect by minus the average score of user rating for the project. If is the common item set that are rated by user I and user j, and are separately the rate which is rate by user I and j, then the similarity between user I and user j is computed used Formula 3.

$$sim(i, j) = \frac{\sum_{c \in I_{i,j}} (R_{i,c} - \overline{R}_i)(R_{j,c} - \overline{R}_j)}{\sqrt{\sum_{c \in I_i} (R_{i,c} - \overline{R}_i)^2} \sqrt{\sum_{c \in I_j} (R_{j,c} - \overline{R}_j)^2}} \quad (3)$$

Where $R_{i,c}$ is the rate of item c by user I; $\overline{R}_i$ and $\overline{R}_j$ are respectively the average rate for the whole items by user I and user j.

D. Pearson Correlation

If the common item set is Iij which include the items rated by user i and user j, then $sim(i, j)$ of the Pearson correlation similarity of two users I and j is defined as Formula 4.

$$sim(i, j) = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \overline{R}_i)(R_{j,c} - \overline{R}_j)}{\sqrt{\sum_{c \in I_{ij}} (R_{i,c} - \overline{R}_i)^2} \sqrt{\sum_{c \in I_{ij}} (R_{j,c} - \overline{R}_j)^2}} \quad (4)$$

Where $R_{i,c}$ is the rate of item c by user I; $R_{j,c}$ is the rate of item c by user j. $\overline{R}_i$ and $\overline{R}_j$ are average value of rate for the whole items by both user x and user y.

III. OUR APPROACH

Collaborative Filtering is an important estimation technique in the information retrieval research domain, rational selection courses are an important content to improve students' confidence and teaching quality. Collaborative Filtering is used in course recommendation system whose process is Fig.1.

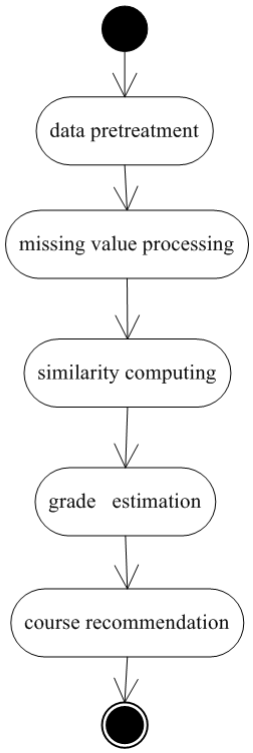

Figure 1. the Process of Course Recommendation based on Collaborative Filtering

The following definitions are given firstly for the convenience of description.

A grade table with m students and n course is described as a m ×n students - curriculum matrix whose rows of matrix are students, and columns of matrix are courses.

$s_i \in \{s_1, s_2 \cdots s_m\}$ denote i-th student,

$c_j \in \{c_1, c_2 \cdots c_n\}$ denote j-th course, and

$g_{ij} \in \{g_{11}, g_{12} \cdots g_{mn}\}$ denotes grade of course c_j for student s_i.

A. Data Pretreatment

Since each metric has different value range, this first step normalizes values of metrics so that the value range becomes [0, 1]. There are non-quantitative values in the set of attributes of projects, such as Boolean, numeric, so non-quantitative values should be processed at first. The 3 steps are used to pretreat grades of these students.

- The student ID will be transformed into a sequence from the beginning of S1, the curriculum will be transformed into a sequence from the beginning of K1.
- If the score g_{ij} is for numeric, then formula (5) is used.

$$nor(g_{ij}) = \frac{g_{ij} - \min(G_j)}{\max(G_j) - \min(G_j)} \quad (5)$$

In that, G_j is range of attribute j for the whole students, $\max(G_j)$ and $\min(G_j)$ denote the maximum and the minimum in the grade table of course j.

- If g_{ij} is for fuzzy value, then two steps are used to normalize. Firstly, the fuzzy value is converted to number start from 1 based on the level from low to high. Secondly, the method for numeric is used.

B. Missing Value Processing

- One of the practical problems in using the estimation methods is that the historical grade database usually contains substantial numbers of missing values. MDTs can give bad influences to the accuracy of estimation, so some complementary techniques have been developed for dealing with missing values. The techniques were: list wise deletion, mean imputation and some types of hot-deck imputation [14]. List wise deletion is the simplest technique to ignore data sets that have missing values. Mean imputation is a technique to fill the missing values on a variable with the mean of data sets that are not missing. Hot-deck imputation is alternative forms of imputation that are based on estimates of the missing values using other variables from the subset of the data that have no missing values. As there is a large of data in grade database, these students having missing scores are deleted directly in the paper.

C. Similarity Computing

In this step, similarity sim (pa, pi) is computed between the target project pa and other projects pi. Many algorithms have been proposed to compute sim (pa, pi) in CF [9-11].

The similarity between two documents is often evaluated by treating each document as a vector of word frequencies and computing the cosine of the angle formed by the two frequency vectors. The method to compute sim (pa, pi) is used in the paper, where projects take the role of documents, attributes take the role of words, and values of the attributes take the role of word frequencies. The sim (pa, pi) is defined between the target project pa and other projects pi as formula (6):

$$sim(s_a, s_i) = \frac{\sum_{j \in S_a \cap S_i} nor(g_{aj}) \times nor(g_{ij})}{\sqrt{\sum_{j \in S_a \cap S_i} \left(nor(g_{aj})^2\right)} \sqrt{\sum_{j \in S_a \cap S_j} \left(nor(g_{ij})^2\right)}} \quad (6)$$

In that, S_a and S_i denote separately grade set of the student S_a and S_i, the $sim(S_a, S_i)$ is the similarity of S_a and S_i whose value range is [0 , 1].

D. Grade Estimation

A grade is estimated for the target student S_a after $sim(s_a, s_i)$ is computed. The steps are as following: Firstly, the k-nearest students are chosen based on similarity. Then the weighted sum is employed to compute grade whose value is computed as the sum of the metrics' values given by the other students similar to S_a. Each value is weighted by the corresponding the $sim(s_a, s_i)$ between S_a and S_i. Formally, the value is defined using formula (7).

$$\hat{g}_{ab} = \frac{\sum_{i \in k-nearest} g_{ib} \times sim(S_a, S_i)}{\sum_{i \in k-nearest} sim(s_a, s_i)} \quad (7)$$

Where k-nearest students denotes set of k students chosen (called neighborhoods) that have highest similarity with S_a.

E. Course Recommendation

According to the estimated value of the grade，8 courses are recommended with the top grades.

IV. EXAMPLE

A. Score Estimation based on Collaborative Filtering

As an example, some grades of students of a class in specialized in computer for 1 year are taken to show the method is feasible in score prediction. These scores are processed using formula (1), and some of the results are shown in Table 1:

TABLE I THE GRADE TABLE PRETREATED

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
S1	0.613	0.619	0.857	0.400	0.542	0.733	0.793	0.550	0.750	0.761
S2	0.806	0.476	0.000	0.350	0.627	0.500	0.586	0.400	0.583	0.775
S3	0.613	0.381	0.571	0.600	0.356	0.367	0.828	0.450	0.500	0.634
S4	0.323	0.238	0.714	0.050	1.000	0.933	0.759	1.000	0.750	0.113
S5	0.516	0.143	0.714	0.525	0.847	0.567	0.207	0.550	0.417	0.775
S6	0.968	0.238	0.286	0.675	0.695	0.367	0.931	0.250	0.750	0.775
S7	0.903	0.286	0.286	0.425	0.424	0.800	0.483	0.250	0.917	0.775
S8	0.613	0.905	0.714	0.700	0.898	0.667	0.897	0.350	0.833	0.845
S9	0.516	1.000	0.857	0.425	0.695	0.067	0.310	0.400	0.667	0.718
S10	0.774	0.476	0.714	0.850	0.593	0.700	0.966	0.650	0.833	0.831
S11	0.839	0.619	1.000	0.825	0.695	0.467	0.690	0.450	0.000	0.831
S12	0.516	0.333	0.571	0.000	0.559	0.033	0.828	0.300	0.250	0.127
S13	0.516	0.476	0.571	0.400	0.797	0.433	0.759	0.400	0.583	0.620
S14	0.774	0.190	0.714	0.625	0.373	0.000	0.690	0.050	0.667	0.648
S15	0.935	0.238	0.857	0.825	0.678	0.500	0.862	0.600	0.750	0.817

The similarity between the students is computed, the 10 maximum value of the similarity are look as the nearest neighbor to estimate score, and the results are shown in Figure 2:

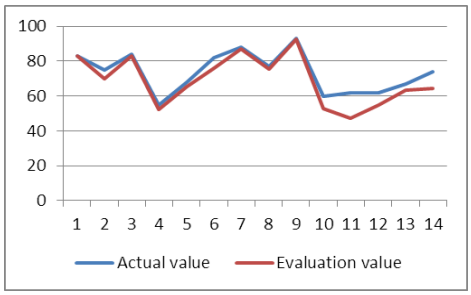

Figure 2. the Result of Score Estimation

B. *Result Evaluation*

The mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes [14], so it is used to measure forecast error in the paper. MAE is given by formula （8）

$$MAE = \frac{\sum_{i=1}^{n} \left| g_i - \hat{g}_i \right|}{n} \quad (8)$$

In that, g_i is the actual value of course I, $\hat{g}_i$ is the evaluation value of course I, and n is the number of courses evaluated.

The MAE is computed using the data of Fig 2, and the result is 4.572104 that shows the deviations between predicted and actual values are within 5.

V. CONCLUSIONS

Collaborative Filtering is an important estimation technique in the information retrieval research domain. It has been successfully applied in both information filtering and E-commerce applications. CF is applied in grade estimation in the paper, which chooses similarity students by similarity, and then the grades of the student are predicted using weight sum, therefore, courses are recommended based on scores. The results of experiment show the deviations between predicted and actual values are within 5. The missing values don't be considered in the paper, so the method to deal with missing value is a future research direction.

REFERENCES

[1] Qi Youran,Pan Zhieheng , Luo Jing. The Mathematical Model of the University Course Recommendation System[J]. Acta Scientiarum Naturalium Universitatis ankaiensis.2011.8:50-52

[2] JIANG Xinlai. Practice and Research of Selecting Courses in L ight ofUniversity Credit System[J]. Journal of Changzhou Institute of Technology.2010.12:87-89

[3] Liu Huihui. The practice and enlightenment of the total quality management of United States[J]. Journal of Jiamusi College of Education.2013.10:190-192

[4] Zhou Lijuan, Xu Mingsheng, Zhang Yanyan.Model of recommended courses based on collaborative filtering[J]. ApplicationR esearch ofComputers.2010.4:1315-1318

[5] Pan Wei.Research on Personalized Courses Recommendation System Based on Collaborative Filtering Technology[J]. Journal of Modern Information.2009.5:193-195

[6] Ralph Bergmann. Introduction to case-based reasoning.http://www.dfki.uni- kl.de/~aabecker/Mosbach/Bergmann-CBR-Survey.pdf,2014.12

[7] Watson. Case-based reasoning is a methodology not a technology. knowledge-based system. elsevier,1999:304-307.

[8] Collaborative filtering[EB/OL]. https://en.wikipedia.org/wiki/Collaborative_filtering.2015.5

[9] Benjamin Marlin. Collaborative Filtering: A Machine Learning Perspective[EB/OL]. http://wenku.baidu.com/link?url=kWXBzZrakdvWc0Yl132N6hZZvRmrmKQEa1z6yAlUPCacrXxPyxQZ0JASF0uyoUW35FTvPEm6XENh1Ra9kj5enbNZqKSPCL-itdfo3yueskm.2015.3

[10] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-Based Collaborative Filtering Recommendation Algorithms.ACM,2001:286-289.

[11] Francesco Ricci and Lior Rokach and Bracha Shapira, Introduction to Recommender Systems Handbook, Recommender Systems Handbook, Springer, 2011:1-35

[12] Euclidean distance, https://en.wikipedia.org/wiki/Euclidean_distance，2015.8

[13] Pearson Correlation Coefficients, http://baike.baidu.com/link?url=RfN3xBwLjuYgg6BFT2cQEKWHADgy_KUVjObXC0Kd6CcdOxTVF2hKT-scdPwjtK1UXYIYEcATIehBYsT3MP6hJa,2015.3

[14] Mean absolute error[EB/OL]. https://en.wikipedia.org/wiki/Mean_absolute_error.2015.5