# Research on primary school Chinese character recommendation algorithm based on collaborative filtering

# Research on primary school Chinese character recommendation algorithm based on collaborative filtering

**Tongtong Zhang[1, a], Yan Yang[2, b], Yucong Duan[3, *]**

[1]Hainan University College of Information Science & Technology Hainan University Haikou, China
[2]Xi'an Deyatong Technology Co., Ltd. R & D department Xi’an, China
[3]Hainan University College of Information Science & Technology Hainan University Haikou, China

*Corresponding author e-mail: duanyucong@hotmail.com, [a]407594351@qq.com, [b]277160299@qq.com

**Abstract**. Collaborative filtering is to calculate the similarity between users by taking all the preferences of a user as a reference value. After finding adjacent users, the information mining of relevant recommendations can be realized through similar behavior history. This paper proposes a Chinese character recommendation algorithm based on user collaborative filtering technology by digging deeply into the recorded information of students' Chinese character learning.

## 1. Introduction

The application of modern information technology to improve the literacy quality of the lower grades of primary schools is widely used in the current classroom [1], which has a great impact on students' vision and sense organs and can attract students' attention. In the classroom, the use of sound, animation, pictures and other forms, can mobilize the enthusiasm of students, active classroom atmosphere, students can also accept this teaching model. Therefore, experts and teachers began to focus on modern information technology focus on the use of multimedia technology to improve the efficiency of primary school literacy in the lower grades. In general, there are many research achievements in improving literacy teaching by using modern information technology in China., but most of the research is all about the study of the theory of the modern information technology teaching [2], few studies have been done on the application of modern information technology in the process of Chinese literacy teaching, the specific forms and strategies of modern information technology assisted literacy teaching. Therefore, this paper proposes a Chinese character recommendation algorithm based on user collaborative filtering technology to realize personalized service of Chinese character learning.

## 2. Collaborative filtering

Collaborative filtering is simply to use the preferences of a group of like-minded people with common experience to recommend information that users are interested in. At present, there are two methods to realize collaborative filtering: one is user-based collaborative filtering algorithm, and the other is item-based collaborative filtering algorithm.

### *2.1. User-based Collaborative Filtering*

This paper focuses on the research of user collaborative filtering algorithm, which relies on the analysis of a large number of users' historical behavior data, and then excavates the items or content they may like. For example, if both user A and B have purchased three books, namely, *Green's Fairy Tales*, *Andersen's Fairy Tales* and *One Thousand and One Nights*, and scored five points for all three books, then we can infer that A and B belong to the same category of users, and then we can recommend the book X read by user A to user B. User collaborative filtering algorithm mainly relies on a large number of users' historical behavior score data, analyzes users' preferences for items to find a list of similar neighbor users, and then makes results recommendation based on the historical behavior of these similar users [3]. That is to say, the user collaborative filtering algorithm is to recommend to the target user what is of interest to the most similar users.

### *2.2. Similarity calculation formula introduction*

To calculate the degree of similarity between users and select an appropriate similarity calculation model, this paper chooses cosine similarity and the Jaccard similarity is used as the calculation factor.

Cosine similarity algorithm is a common method to calculate similarity. Cosine similarity, also known as cosine similarity, is to evaluate the similarity of two vectors by calculating their angle cosine values. The closer the cosine value is to 1, the angle between the two vectors is 0, which means the more similar the two vectors are. This is called cosine similarity. The calculation formula is as follows:

$$sim(u,v)^{cos} = \frac{\sum_{i\in I} r_{ui} * r_{vi}}{\sqrt{\sum_{i\in I} r_{ui}^2}\sqrt{\sum_{i\in I} r_{vi}^2}}$$

Jaccard similarity coefficient is mainly used to compare the similarity and difference between limited samples. The calculation method is as follows:

$$sim(u,v)^{Jaccard} = \frac{|\, \mathrm{I}_{U} \cap \mathrm{I}_{v}|}{|\, \mathrm{I}_{U} \cup \mathrm{I}_{v}|}$$

## 3. Process based on user collaborative filtering

There are three main processes of user-based collaborative filtering algorithm. First, the user model is established; second, the nearest user is found; and finally, the list of recommended items is generated.

### *3.1. Build the "user-Chinese character" model*

The correct rate of each student's Chinese character is expressed by dividing the correct number of times of each student's writing by the sum of times of writing, which is used to measure the degree of the student's mastery of the Chinese character.

**Table 1.** Examples Of "Student-Chinese Character" Correctness Matrix

| | tu | hua | long | ran | …… |
|---|---|---|---|---|---|
| student1 | 66% | 54% | 95% | 60% | …… |
| student2 | 50% | 63% | 41% | 48% | …… |
| student3 | 48% | 87% | 86% | 65% | …… |
| student4 | 88% | 68% | 79% | 66% | …… |
| …… | …… | …… | …… | …… | …… |

### *3.2. Find Similar Users*

Cosine similarity is often calculated for users with low similarity to obtain high similarity values, such as (4, 3), (2, 1), the two preferences may be opposite, but the similarity value is calculated very high.

*3.2.1. Invalid data filtering.* In data processing, we will introduce the threshold of "Chinese character mastery rate" to further mine the data in the "student-Chinese character" accuracy rate matrix to screen out reasonable and effective data. "The control rate of Chinese characters" the threshold as the judging criteria system to determine whether a student to master Chinese characters, as a student of a Chinese character is greater than or equal to the correct valve value that the students to master the Chinese characters, so accuracy is less than the threshold for Chinese characters, represents not mastering the Chinese characters, when calculating the abandoned, can improve the calculation accuracy of cosine similarity.

On the basis of common sense we know different its ease of Chinese characters is different, we use the $\xi$ to measure to the difficulty of the Chinese characters, and use x to represent the number of strokes of Chinese characters. In the general case for a Chinese character, the more the number of strokes, then the Chinese character more complex and more difficult to master, is deduced and the inversely proportional relationship between the average number of strokes of Chinese characters is known as 7, the image passes through (1, 1) and (7, $\xi$) , the image is as follows:

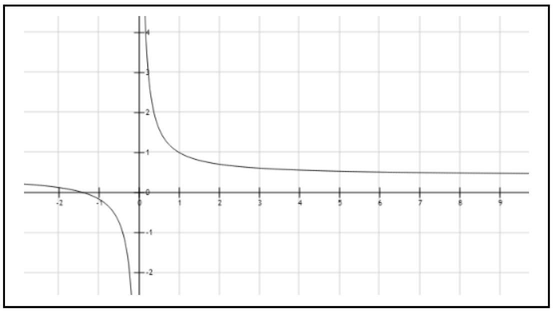

**Figure 1.** Function Images of $\xi$ and *x*

The relationship between $\xi$ and *x* can be expressed as follows:

$\xi=\frac{5x+7}{12x}$ $x \in (0,+\infty)$

If $\lambda$ is used to represent the mastery rate of Chinese characters, the more strokes the Chinese characters have, that is, the greater the value of $\xi$, the smaller the value of $\lambda$. Its image is shown in the following figure:

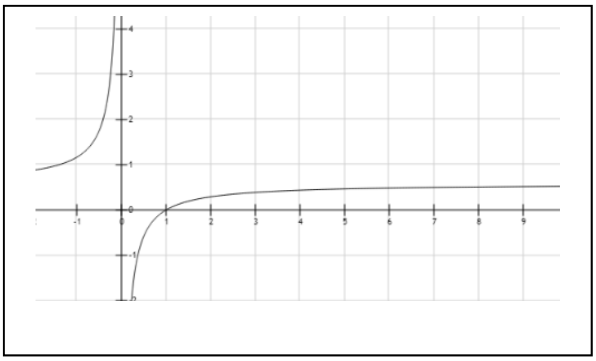

**Figure 2**. Function Images of $\xi$ and $\lambda$

Then the functional relationship between $\xi$ and $\lambda$ is:

$\lambda=1-\frac{(7\xi-1)(x-1)+6}{6x}$ $x \in (0,+\infty)$

Before calculating the similarity officially, the data with the correct rate less than $\lambda$ should be filtered first. For example, assuming that $\xi$ is 0.5, after processing Table 2, a new "Student-Chinese Character" correct rate matrix can be obtained as follows:

**Table 2.** Examples of "Student-Chinese Character" correct rate matrix

| | tu | hua | long | ran | …… |
|---|---|---|---|---|---|
| student1 | 66% | 54% | 95% | 60% | …… |
| student2 | 0 | 63% | 0 | 48% | …… |
| student3 | 0 | 87% | 86% | 65% | …… |
| student4 | 88% | 68% | 79% | 66% | …… |
| …… | …… | …… | …… | …… | …… |

Then the cosine similarity formula is introduced to calculate the similarity between each user.

*3.2.2. Find Similar Users.* After calculating the similarity between users, it is necessary to generate a set of adjacent users. In this paper, the traditional Top-N and threshold filtering are abandoned when selecting adjacent users. The difference ξ between the similarity and the true similarity is estimated by Hoeffding's boundary inequality [4]. Select similar values that match the extreme case and add them to the top-k sequence as adjacent sets of users.

Assuming that there is a benchmark user $u_*$, when the user's neighbor is selected, for any user $u_x$EU, if sim (u, $u_x$)> sim(u, $u_*$) is satisfied, $u_*$is the neighbor of user U. The benchmark user $u_*$here is the virtual user, and it is assumed that the Chinese character user u learned is the same as the Chinese character learned by user $u_*$, and its accuracy rate is the average of the accuracy rate of other users calculated by comparing with user u.

When determining whether u2 is a close neighbor of u1, if u2 is a close neighbor of u1, sim($u_1$, $u_2$)> sim($u_1$, $u_x$)must be satisfied according to definition 1. In order to be true in any case, the inequality is true in the extreme case, that is, when sim ($u_1$, $u_2$) takes the minimum value and sim($u_1$, $u_x$)takes the maximum value.

$$\overline{sim(u_1,u_2)} - \varepsilon_1 \geq \overline{sim(u_1,u_*)} + \varepsilon_2$$
$$\overline{sim(u_1,u_2)} - \overline{sim(u_1,u_*)} \geq \varepsilon_1 + \varepsilon_2$$

$\varepsilon_1$ is the Hoeffding boundary value of $\overline{sim\ (u_1,\ u_2)}$, and the calculation formula is:

$$\varepsilon_1 = \sqrt{\frac{R_{(u_1,u_2)}{}^2 .\ln\ (\frac{1}{\delta_{(u_1,u_2)}})}{2n}}$$

Among them, $R_{(u_1,u_2)}$is the similarity of $(u_1,u_2)$the value of $\delta_{(u_1,u_2)}$ is $sim(u_1,u_2)^{Jaccard}$ and N is the set of Chinese characters when calculating $(u_1,u_2)$similarity.

*3.3. Generate Chinese Character Recommendation List*

When these nearest neighbors are calculated, the data of these neighbors are combined to predict the correct rate of recommending target users in learning new Chinese characters, generate the prediction of the target users to be recommended [5], and then recommend the items with the highest prediction value to users first. Let $S_{ua}$ denote the nearest neighbor set of target user a, p(a, j) denotes user a's predictive rating of Chinese character J. It is mainly obtained by calculating the accuracy of Chinese characters in the nearest neighbor set $S_{ua}$ by user a. The calculation method is as follows:

$$p(a,j) = \overline{R_a} + \frac{\sum_{n \in S_{ua}} sim(a,n) * (R_{n,j} - \overline{R_n})}{\sum_{n \in S_{ua}} |sim(a,n)|}$$

sim(a, n) Represents the similarity between user a and user n. $R_{n,j}$ denotes user n's evaluation of Chinese character J. $\overline{R_a}$ and $\overline{R_n}$ represent the average correct rate of learning Chinese characters for users a and n, respectively.

## 4. Experimental Analysis

The development platform of the system is Eclipse, the database management system is SQL Server 2005, and the programming language of the system is JAVA.

### *4.1. Build a Two-Dimensional Matrix Model of Student-Chinese Characters*

The user-item two-dimensional matrix is input into the system, and the similarity between users is calculated by the similarity formula. The set of user's personality preferences Item is obtained. Finally, the recommendation list of each user is obtained. In SQL Server 2005, the word_stat database of student information is constructed, which is used to save various data lists in recommendation system.

### *4.2. Generate Chinese characters recommendation*

In order to find the user set that is similar to the target user, this paper estimates the difference between the similarity value and the true similarity value by using Hoeffding boundary inequality, calculates the similarity between the target user and other users, and adds the value that meets the Hoeffding boundary inequality condition to the list of similar user set.

This paper systematically analyses the correct rate of Chinese characters that similar users of current users have learned, predicts the correct rate of Chinese characters that current users will learn, makes personalized Chinese character recommendation for each user, and recommends Chinese characters that neighbor users have learned and are suitable for current users to current users.

### *4.3. Result analysis*

**Table 3.** Results of Chinese characters recommendation

| student | Similar student | Recommendation and accuracy rate |
|---|---|---|
| student1 | 2, 4, 6 | ni(0.87), de(0.57), lian(0.35)··· |
| student2 | 1, 7, 6, 9 | ni(0.69), liang(0.68), li(0.59)… |
| student3 | 8 | kai(1.06), you(0.99), ni(0.91)… |
| …… | …… | …… |

From Table 3, it can be seen that the user-based collaborative filtering algorithm successfully predicts and recommends easy-to-learn Chinese characters to users. The advantage of this algorithm is to analyze the similarity of learning degree between users in complex Chinese character learning field for reasonable Chinese character recommendation, avoiding the analysis of learning content, and has a strong ability to recommend Chinese characters. Timeliness. However, due to the small amount of data, there are still errors in each user's prediction. Further data analysis by expanding the user base or adding Chinese character structure factors has more obvious effect.

## References

[1] J. SU Danrui, LI Yubin, Review of Domestic M-learning in Recent Five Years, Chinese Educational Technology &Equipment, 2016, pp.8 - 10.

[2] D. Da Ren, A Study on the Applied Status and Improving Suggestions of Learning Chinese Characters on Mobile Apps, Shanghai Normal University, 2014, pp.6 - 7.

[3] D. Lang Peng, Research and Implementation of User-based Collaborative Filtering Recommendation Technology, Ningxia University, 2014.

[4] J.Wang Hongliang,Zhao Li, Hoeffding Clonal Selection Algorithm and Its Application in Associative Classification, Journal of Chinese Information Pro-cessing, 2012,pp.65 - 72.

[5] J. Ding Shaoheng, Ji Donghong,Wang Lulu, Collaborative filtering recommendation algorithm based on user attributes and acores, Computer Engineering and Design, 2015.