# Analysis of data-driven computer course system

ZHANG Xin[1], ZHANG Shaohong[1*], LIU Jinlong[2]

[1] Guangzhou University, School of Computer Science and Cyber Engineering, Guangzhou Guangdong 51064，China

[2] School of software, Shandong University, Jinan Shandong 250101, China

**Abstract.** The development of computer technology has brought great convenience to the production and life of the society, but also led to the rapid growth of the demand for computer talents. Whether we can train more and better computer talents has become an inevitable development of society. In the process of training computer talents, the cultivation of undergraduates is very important. In order to analyze the cultivation process of undergraduates systematically, this paper selects some undergraduate teaching outline documents of computer science and technology major in some universities for analysis. After using TF-IDF to quantify the curriculum information in the syllabus, this paper proposes to use multiple K-means clustering algorithm to calculate the similarity between courses; the similarity matrix between schools is calculated by using the method of sum and average of the maximum similarity without repetition, and the similarity between schools is analyzed by using the calculation results. The work of this paper will provide data reference for computer students, teachers and curriculum.

**Keywords:** Computer Science and Technology; TF-IDF; multiple K-means clustering; maximum similarity

## 1 Introduction

According to the statistics of the Ministry of industry and information technology, the operating revenue of the computer industry in 2017 was 468.306 billion yuan, a year-on-year increase of 28.02%, an increase of 11.84% compared with the 16.19% growth rate of the revenue in 2016, and the rising trend of the computer industry continues. In Colleges and universities, whether we can make feasible teaching plans and teaching arrangements is the basis of whether we can cultivate computer talents that meet the needs of the times. Many scholars are also studying on this [1]. This paper studies the existing universities that have already opened computer courses, and selects the undergraduate syllabus of computer science and technology major in some universities as the original data and conducts analysis and research.

In this paper, in the specific implementation process, first of all, the curriculum information is extracted from the syllabus text, the extracted curriculum description information is segmented, and the segmentation results are vectorized. TF-IDF algorithm is used to calculate the weight of the segmentation results and realize the text

* Corresponding author: Zhang Shaohong, zimzsh@qq.com

vectorization. Then, the similarity between different courses is calculated, which is used in this paper a variety of algorithms are calculated and compared, including cosine distance and Tanimoto coefficient algorithm. According to the characteristics of the purpose of this study, the method of multiple clustering and comprehensive analysis is proposed to get more accurate similarity value between courses and courses. Finally, the similarity matrix between different schools is used to calculate the similarity between schools. Three different methods are used in the calculation, namely, the overall similarity matrix of the inter-school curriculum is summed and averaged, the rows and columns in the similarity matrix find a maximum value and the average, and the rows and columns are not repeatable select the rank value and sum the average, the calculated value can represent the similarity between schools.

## 2 Quantitative representation of curriculum

In order to convert the text data of the original course into data that can be processed by the computer, a series of data processing operations are required. First, standardize the text format of the syllabus file; second, extract the processed text to obtain the description information of all the courses in the school; third, segment the course description information through the tokenizer; finally, the TF-IDF algorithm [2] is used to weight and quantize the segmentation results of the course description information, as shown in Figure 1.

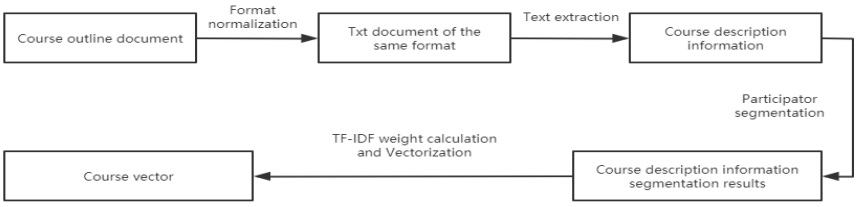


**Fig. 1.** Flow chart of course vectorization processing

It can be seen from the content of flow chart that how to turn the segmentation result of course description information into course vector is the key of course vector processing. In the research process of this paper, in order to transform the result of course description information segmentation into course vector, several mainstream algorithms are considered, including Textrank algorithm [3], Term Frequency, Term Frequency–Inverse Document Frequency. In the selection process of these mainstream algorithms, the characteristics of each algorithm and the results of this study are considered. In order to get better results of school similarity, TF-IDF algorithm is used to calculate the weight of the segmentation results, and the calculated results are transformed into the vector information of the course.

In order to facilitate the implementation of the algorithm, the samples of course documents are recorded as $\mathrm{D}=\{D_i|i=1,2,\cdots,m\}$, where m means there are m course documents in total. Each course file $D_i$ contains the description information of the

course, and the segmentation results are obtained after segmentation by the word breaker. All words are recorded as $W=\{W_i|i=1,2,\cdots,n\}$, where n represents that there are n different words in all course documents [4]. Using TF-IDF algorithm to vectorize the course description file after word segmentation, the specific implementation method is as follows:

(1) Calculating the inverse document frequency index of words

The IDF of a word indicates the ability of the word to distinguish text in the sample. The larger the IDF index of a word, the greater the ability of the word to distinguish text. The formula $W_i$ for calculating the $IDF_i$ of words is defined as follows:

$$IDF_i=\ln\frac{|D|}{\left|\{j:\ W_i\in D_j\}\right|}$$

The numerator $|D|$ represents the total number of all files in the sample. The denominator $\left|\{j:\ W_i\in D_j\}\right|$ represent the number of course files containing word $W_i$. According to the formula, we can get the inverse document frequency index of a word in the sample file [5]

(2) Calculate word frequency of words in each file

For a course file, the word frequency of the word $W_i$ indicates the frequency of the word in the course file.

$$TF_{k,\ i}=\frac{n_{k,\ i}}{\sum_j n_{k,\ j}}$$

The numerator $n_{k,\ i}$ indicates the number of word $W_i$ appears in course file $D_k$. The denominator $\sum_j n_{k,\ j}$ indicates the number of words $W_i$ appear in all course file.

(3) Calculate TF-IDF weight value of course file

The TF-IDF weight value of word $W_i$ in course file $D_k$ is defined as follows:

$$TF-IDF=TF_{k,i}*IDF_i$$

It can be seen from the formula that the TF-IDF weight value of the word $W_i$ in the course document $D_k$ is directly proportional to the word frequency and the inverse document frequency index of the word. Through the calculation of the formula, the weight value of different words in each course file can be obtained. The weight value of all words in a course file constitutes the course vector of the file. The course description information will become a vector with n-dimension, and there are m such course vectors, which all of them are recorded as $X=\{X_i|i=1,2,\cdots,m\}$.

## 3 Curriculum similarity calculation

After the vectorization of course description information, each course becomes an n-dimensional vector $X_i$. In order to calculate the similarity between different courses, three similarity calculation methods are selected in this paper. Cosine distance and Tanimoto coefficient (generalized Jaccard coefficient) are two similarity calculation

methods, which are often used for similarity calculation of text vector. In addition, this paper selects clustering algorithm to calculate the similarity between different courses, and uses the algorithm principle of K-means clustering algorithm. The course vector is recorded as $X_i = \left(x_{i,1}, x_{i,2}, \cdots, x_{i,n}\right)$.

### 3.1 Cosine distance

Cosine distance between two course vectors refers to the cosine value of the angle between two course vectors, which is compared with Euclidean distance, Manhattan distance, Chebyshev Distance and other commonly used similarity calculation methods, cosine distance calculation results are non-negative values, and the distribution range of the values is [0,1], the size of the numerical results is proportional to the similarity, cosine distance calculation results, more in line with the requirements of this experiment.

The principle of cosine distance is to calculate the cosine value of the angle between two vectors, and use the cosine value to measure the similarity between vectors. In two-dimensional space, when two vectors are in the same direction, the cosine value of the angle between vectors is 1. When two vectors are perpendicular to each other, the cosine value is 0. Cosine distance is used to measure the similarity between two vectors, which is very suitable for high-dimensional text analysis [6].

In the experimental data, when calculating the cosine value of the angle β between the course description vector $X_1 = \left(x_{1,1}, x_{1,2}, \cdots, x_{1,n}\right)$ and the course description vector $X_2 = \left(x_{2,1}, x_{2,2}, \cdots, x_{2,n}\right)$, the calculation formula is as follows:

$$\cos\beta = \frac{\overrightarrow{X_1} \cdot \overrightarrow{X_2}}{\overrightarrow{X_1} \cdot \overrightarrow{X_2}} = \frac{x_{1,1}x_{2,1} + x_{1,2}x_{2,2} + \cdots + x_{1,n}x_{2,n}}{\sqrt{x_{1,1}^2 + x_{1,2}^2 + \cdots + x_{1,n}^2} + \sqrt{x_{2,1}^2 + x_{2,2}^2 + \cdots + x_{2,n}^2}}$$

After the similarity between courses is calculated by cosine distance formula [7], the similarity between courses is stored in the form of matrix.

### 3.2 Tanimoto coefficient

Tanimoto coefficient is also called generalized Jaccard coefficient. The common calculation of Jaccard coefficient only considers whether the word is included in the course description vector, and does not consider the weight information after the word calculation, so this paper uses Tanimoto coefficient to calculate the similarity between courses. The calculation formula of Jaccard coefficient between course vector and course vector [8] is as follows:

$$Tan_{1,2} = \frac{\overrightarrow{X_1} \cdot \overrightarrow{X_2}}{\overrightarrow{X_1}^2 + \overrightarrow{X_2}^2 - \overrightarrow{X_1} \cdot \overrightarrow{X_2}} = \frac{\sum_{i=1}^{n} x_{1,i} * x_{2,i}}{\sum_{i=1}^{n} x_{1,i}^2 + \sum_{i=1}^{n} x_{2,i}^2 - \sum_{i=1}^{n} x_{1,i} * x_{2,i}}$$

Similar to cosine distance, calculate according to the above formula and put the calculated results into the matrix.

### 3.3 Multiple K-means clustering algorithm

Among all clustering algorithms, K-means clustering algorithm is a classic prototype clustering algorithm. The overall idea of the algorithm is to divide the given data set into K clusters, so as to minimize the square error of clustering results [9]. The

calculated cluster is recorded as $C=\{C_i|i=1,2,\cdots,k\}$, the $\mu_i=\frac{1}{|C_i|}\sum_{d\in C_i} d$ represents the mean vector of the cluster $C_i$, and the greedy algorithm is used to find the cluster division under the minimum square error as following.

$$E=\sum_{i=1}^{k}\sum_{d\in C_i} d-\mu_i{}_2^2$$

In this experiment, we hope to be able to calculate the similarity between the course vectors. In the idea of clustering, it can be understood as the possibility that two course vectors can be divided into the same cluster. Therefore, for the purpose of this experiment, the traditional K-means clustering algorithm is improved, and the K-means clustering calculation is carried out for many times for the course file data set, and the K-values of each time are different, and the clustering results are transformed into the possibility of two course vectors in the same cluster, and the realization process is shown in Figure 2.

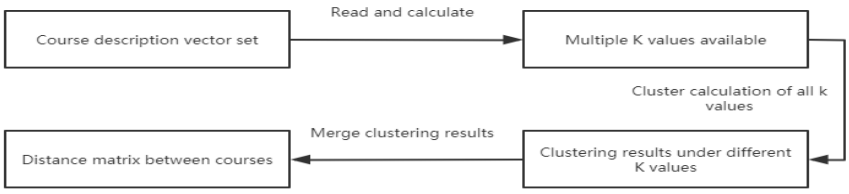


**Fig. 2.** Flow chart of multiple k-means clustering

Firstly, all the course description vectors will be read, and the K value used in clustering operation will be calculated and determined according to the number of vectors. Secondly, use different K values to calculate the K-means, and store all clustering results in a matrix. In the matrix, the vector represented by the horizontal and vertical coordinates is recorded as 1 when it is in the same cluster, and 0 when it is not in a matrix. Finally, all clustering results are combined. In order to make the results more representative, all clustering matrices are added and averaged, and the average results are taken as the similarity between vectors, and the results are saved in the form of matrix.

Intuitively speaking, the larger a numerical result in the calculated result matrix is, the more likely it is to be in the same cluster after clustering, and the smaller the value is, the less likely it is to be in the same cluster. And the value range of calculation results is [0,1], which is convenient for later calculation.

## 4 School similarity calculation

The similarity degree of the opening plan between the two schools is reflected in the number of courses offered by the two schools, the name and content of the courses. How to use the similarity matrix between the two schools to calculate the similarity between the two schools has become the key to calculate the similarity between the

schools. This paper uses three methods to calculate the similarity between schools by using the similarity matrix between courses, and evaluate the three methods.

Firstly, using the similarity matrix between the two schools to calculate the similarity between the two schools, the simplest idea is to sum all elements in the similarity matrix between the courses and calculate the average value, taking the average value as the similarity between schools. Through the description of the average calculation, it can be seen that the smaller the similarity between the two schools after calculation, the more courses with small similarity between the two schools, the higher the curriculum similarity between the two schools; the greater the similarity between the schools, the more courses with large similarity between the two schools, the lower the curriculum similarity between the schools. But for this experiment, the similarity between all courses in two schools is only calculated by simple sum and average operation, and the result value can not directly represent the similarity between schools. If the data is not filtered and all the data is simply summed up and averaged, there will be a lot of irrelevant data affecting the results.

Secondly, in order to avoid the influence of irrelevant data, this paper selects the maximum method to filter the similarity matrix between the two schools' courses, and calculates the filtered data, so as to obtain the similarity between the two schools. The specific implementation process is as follows. Firstly, the similarity matrix between two schools is extracted from the similarity matrix of all courses; secondly, the maximum value of the row and column of the extracted similarity matrix is selected respectively, and the selected maximum value is stored as the data for later operation; finally, the selected data is summed and averaged, and the result value is taken as two Similarity of schools. But even in the same school, the design of different courses is different, that is to say, the course itself is unique.

Thirdly, using this feature, in the process of data screening of course similarity matrix, we should avoid the similarity matching between one course and multiple courses, and take anti repetition measures when screening data, so that the data screened is more in line with this experiment. Compared with the maximum filtering method, the difference is mainly reflected in the process of filtering data in similarity matrix, as shown in Figure 3, the data in the circle in the figure represents the filtering results. When the data is filtered by the maximum value, each row only needs to select the maximum value of the current row. When the data is filtered by the maximum value without repetition, the data columns selected by the previous row cannot appear in the subsequent row data repeatedly.

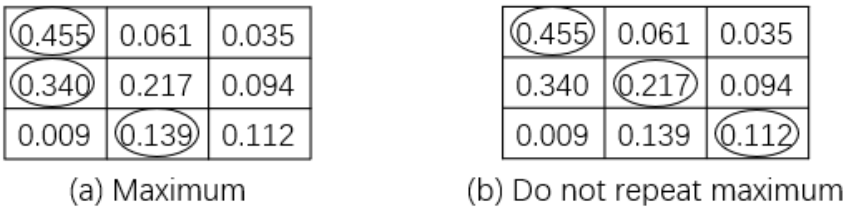


**Fig3** Results of selecting data used two ways

## 5 Experimental results and analysis

In this experiment, three methods are used to calculate the similarity between the course vectors, namely cosine distance, Tanimoto coefficient and multiple K-means clustering algorithm. Taking two courses of Hubei Institute of technology, C language programming course and data structure course as examples, the similarity between the three courses of Southwest University, C language programming course, data structure course and assembly language programming course is calculated in three ways The calculation results are written into the table, as shown in Table 1. In Table 1, C language programming courses and data structure courses of Hubei Institute of technology are used $H_1$ and $H_2$ represented. C language programming courses, data structure courses and assembly language programming courses of Southwest University are used $X_1$, $X_2$ and $X_3$ represented. Among the three similarity calculation methods, the calculation results of multiple K-means clustering are highly differentiated, whether the courses are the same or not, the calculation results differ greatly, with higher differentiation and accuracy.

**Table1** Distance of two university`s course

| Hubei Institute of Technology | Southwest University | Cosine | Tanimoto | Multiple K-means clustering |
|---|---|---|---|---|
| $H_1$ | $\boldsymbol{X}_1$ | 0.7105 | 0.5510 | 1 |
| | $\boldsymbol{X}_2$ | 0.1962 | 0.1088 | 0 |
| | $\boldsymbol{X}_3$ | 0.2227 | 0.1253 | 0 |
| $H_2$ | $\boldsymbol{X}_1$ | 0.0890 | 0.1210 | 0 |
| | $\boldsymbol{X}_2$ | 0.7651 | 0.6196 | 1 |
| | $\boldsymbol{X}_3$ | 0.1283 | 0.0686 | 0 |

Through the above analysis, we can see that the similarity between courses is calculated by using the multiple k-means algorithm, and the similarity matrix between schools is obtained. After the number in the course matrix is filtered by the maximum value of row and column non repetition, the average value is calculated and recorded as the similarity value between schools. The similarity information between schools obtained in this experiment is displayed in the way of thermodynamic diagram, as shown in Figure 4. The color depth of each color block in the diagram indicates the similarity between two schools in the row where the color block is located. From the content shown in Figure 4, it can be seen that the similarity values among universities are evenly distributed between 0-0.6, and the similarity values between Shenyang University of technology and other schools are relatively small, indicating that the similarity between the syllabus arrangement of Shenyang University of technology and other schools is low in the data collected in this experiment.

According to the calculation results of this experiment, the experimental results are analyzed by referring to the ranking of 2018 school comprehensive ability assessment issued by all schools in China Alumni Association, and the ranking of 2016 computer science and technology discipline. In the two lists released, take Qufu Normal Uni-

versity, Shanghai Normal University and Southwest University as examples. The comprehensive strength rankings of the three universities are 201, 111 and 38 respectively. However, in the computer science and technology discipline, the rankings of the three universities are 84, which shows that although there are differences in the comprehensive rankings of the three universities, the evaluation of the computer science and technology discipline The price belongs to the same rank.

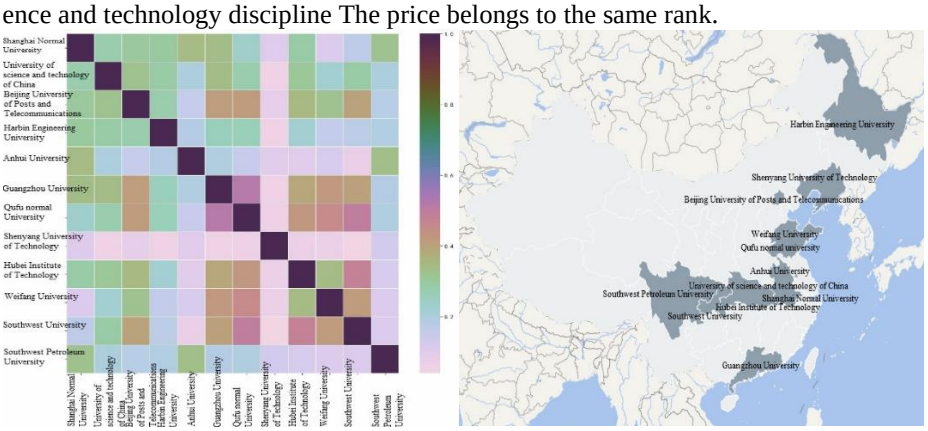


**Fig4** Heat map of school`s distance and School`s location lable

The sample schools used in this experiment and their locations are marked on the map, and the results are shown in Figure 4. It can be observed from the geographical location that the similarity between undergraduate schools in coastal cities is higher, and the similarity between schools in inland areas is higher than that between schools in inland areas and coastal areas. And the similarity between schools is related to the degree of communication between schools. In the same area, schools often exchange teaching experience and curriculum, so the similarity between schools is higher, and the similarity between schools in different areas is smaller.

It can be seen that the maximum similarity between Qufu Normal University and Southwest University is about 0.5. Although the comprehensive ranking of the two schools is quite different, the similarity in the design of professional content is relatively high, which is closely related to the continuous development of computer science and technology in Qufu Normal University; Shanghai Normal University, Qufu Normal University and Southwest University The similarity value of universities is relatively small, about 0.24 and 0.17, respectively. Although the ranking of computer science and technology disciplines in the three schools is the same, there are differences in curriculum construction among the three schools. We can know that the overall level of subject education in schools can be further improved by considering the curriculum, curriculum objectives and arrangements.

## 6 Conclusion

Based on the analysis of the syllabus documents of undergraduate students majoring in computer science and technology in schools of different regions and levels, we

can see that the offering and content setting of teaching courses are closely related to the comprehensive ranking and geographical location of the school. It is very helpful for the school to enhance the communication and study with other schools in the course construction and to improve the strength of computer science and technology major. For a long time, few scholars have analyzed the major from the aspect of curriculum construction, and no one in the undergraduate syllabus of each school has carried out special data sorting. This experiment, from the perspective of the construction of the school's professional curriculum, has been considered and analyzed, and has obtained a very valuable effect. Through the experiment, we hope to cause the school to pay more attention to the construction of the curriculum system Look at.

**Acknowledgment**

The work described in this paper was partially supported by grants from Guangdong Natural Science Foundation of China [Grant No. 2018A030313922], the funding of Guangzhou education scientific research project [Project No. 1201730714], and the Postgraduate Educational Reform project of Guangdong Province [No. 2017JGXM-MS45].

## References

1. Mao Y , Feng Y , Cheng D , et al. Computer curriculum system reform based on system ability training[C]// International Conference on Computer Science & Education. IEEE, 2016.
2. Hui H C, Guangzhou. A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method[J]. Chinese Journal of Computers, 2011, 34(5):856-864.
3. Wu W, Zhang B, Ostendorf M. Automatic Generation of Personalized Annotation Tags for Twitter Users[J].Naacl, 2010:689-692.
4. Xing-Dong S , Ai-Ping L I , Shu-Dong L I . Research and Implementation of Micro-blog Keyword Extraction Method Based on Clustering[J]. Netinfo Security, 2014.
5. Robertson, Stephen. Understanding inverse document frequency: on theoretical arguments for IDF[J]. Journal of Documentation, 2004, 60(5):503-520.
6. Qian G , Sural S , Gu Y , et al. [ACM Press the 2004 ACM symposium - Nicosia, Cyprus (2004.03.14-2004.03.17)] Proceedings of the 2004 ACM symposium on Applied computing, - SAC \"04 - Similarity between Euclidean and cosine angle distance for nearest neighbor queries[J]. 2004:1232.
7. Sohn M W. Distance and cosine measures of niche overlap[J]. Social Networks, 2001, 23(2):141-165.
8. Fligner M A, Verducci J S, Blower P E. A Modification of the Jaccard–Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings[J]. Technometrics, 2002, 44(2):110-119.
9. Kanungo T, Mount D M, Netanyahu N S, et al. An efficient k-means clustering algorithm: analysis and implementation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7):0-892.