International Journal of Database Theory and Application
Vol.10, No.1 (2017), pp.187-196
http://dx.doi.org/10.14257/ijdta.2017.10.1.17

Evaluation Method of College Students' English Proficiency Based on Computer Aided Cluster Analysis

Yanjiao Xiao*

College of Foreign Languages, Southwest Petroleum University, Chengdu 610500, China
**Yanjiao Xiao, nancyxyj@gmail.com*

Abstract

Discovery in databases knowledge means obtain effective, implicit and potentially useful knowledge from a large number data of database. As China's higher education has been transferred to mass education, the scale of school and the number of students is increasing. By using data mining techniques, the author makes the score analysis of National English test (CET-4), mining useful information hidden in the performance data, then provides theoretical basis for the teaching design and management in English teaching. After K value clustering, we can effectively classify the students, so as to carry on the difference teaching, and this classified teaching will improve the quality of English teaching.

Keywords*: Clustering analysis, English learning, knowledge discovery, English score, CET-4*

1. Introduction

Data mining (DM) also known as the database of knowledge found (KDD), it could obtain the effective, implicit, potential and useful knowledge process from a large number of database data [1-2]. It combines the database, data warehouse, statistics, artificial intelligence, machine learning, pattern recognition, information retrieval, domain knowledge is a hot issue in the study of current information technology, is an area in the research of the information technology is very promising [3]. In recent years, in order to adapt to the development of the times and the needs of society, our country's higher education has entered the era of mass education; as China's higher education has been rapid development. In order to adapt to the popularization of higher education and the State encourages schools in various forms of background support gave birth to the independent college has been the rapid development, it has become an important part of higher education in our country, become the important force to promote the process of popularization of higher education in China [4-5]. Students' test scores are not only the most intuitive and effective way to test students' learning effects and teachers' teaching effectiveness, but also an important indicator to measure the quality of education and teaching in a school. Such as the National College CET-4 has been widely recognized by the society, enterprise, CET-4 scores reflect not only students' English learning ability, is also an important indicator to measure the quality of College English teaching, CET-4 scores or even directly affect the success or failure of the job[6]. Traditional student achievement management and analysis is nothing more than a simple sorting and analyzing the results, this can only get some surface information, not to make more in-depth analysis of the relationship between various factors in the teaching process and students' scores and CET-4 scores and other courses. In order to solve this problem, data mining technology is introduced to the analysis of the achievements of students, through on student achievement data mining analysis, from identify and factors influencing students' learning potential of a relationship, reuse analysis the result to find out the daily

ISSN: 2005-4270 IJDTA

activities of education and teaching process in which factors on student achievement has important influence, provide decision-making reference and support for educational and teaching activities, so as to the development of teaching reform, curriculum and management mode reform, thus to improve the quality of education and teaching.

At present, China's higher education has been transferred from elite education to mass education, the scale of running school, the number of students in school. After many years of use of comprehensive academic information management system, the university has accumulated a large number of student data [7]. The data mining technique is applied to the students of the national university English four level examinations (CET-4) performance analysis, dig out the useful information hidden in the data, provide the basis for the improvement of teaching design and teaching management. Established to predict whether students through the CET-4 decision tree, using the ex post pruning method to build the decision tree pruning withdraw classification rules, according to the classification rules to establish college CET-4 achievement prediction model based on C4.5 algorithm, the classification rules of each influencing factor analysis, analysis of the impact of the various factors of CET-4 scores, in order to improve English teaching design and teaching management provide scientific basis, to improve the students English and CET-4 exam pass rate [8]. The K algorithm is used to cluster analysis of the variant K-Modes algorithm on the CET-4 performance of the qualified and unqualified persons. Aiming at the shortcoming of difficult to determine the value of K is a determined K value method, the method first choice relatively large K values using cluster analysis, to calculate the distance between each cluster after cluster, if the distance is below a certain threshold that two clustering is too similar to the merger [9]. After the merger in accordance with the new K value to cluster. After clustering, the characteristics of each cluster are analyzed, and the classification rules obtained by the decision tree model are verified. Can according to the characteristics of different clusters of students for placement of classified teaching, so do teach students in accordance of their aptitude, teaching students according to their aptitude, improve the quality of teaching, improve students' CET-4 exam pass rate.

2. Data Mining Method

Data mining is a combination of database, artificial intelligence, statistics, machine learning, neural network, mathematical statistics and other interdisciplinary subjects. Since the date of birth, the definition of data mining has been given a lot of kinds, simply speaking, data mining is to extract or "dig" knowledge from a large amount of data. At present has been generally accepted definition: data mining is from the large, complete, noisy, fuzzy, uncertain data, extraction of implicit in which people do not know in advance, reliable and useful knowledge. The above definition is defined from the point of view of technology, also has research scholars from the application point of view to define data mining: data mining as the information in the field of a new processing technique for analysis, it mainly in the face of the object is for business, from a large amount of business data extraction calibration data, and then collate the target data. Finally, on the target data were analysis and mining, found hidden in the meaningful business model and rules, then carries on the modeling process of the business model and rules.

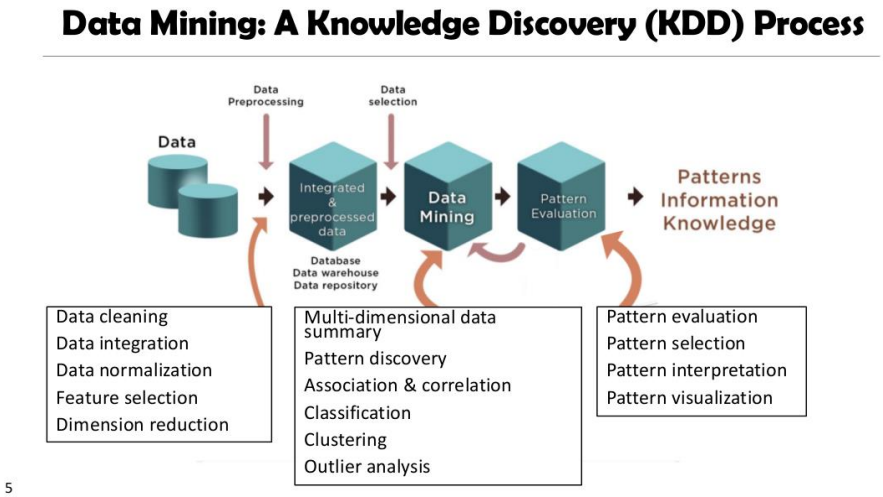


Figure 1. Knowledge Discovery Process

- The data preprocessing stage: data preprocessing is the data source for further processing, to data in the susceptible to noise in the data smoothing, spatial values to fill a vacancy, eliminate the "dirty" data. In the process of data mining, most of the time and energy are used in the data preprocessing stage, the data pretreatment to do good and bad directly affect the quality and accuracy of data mining. Data cleaning is to identify or remove isolated points, the source data is not complete, inconsistent, noise data to fill the default, incomplete data and delete duplicate data. Data integration is the organic integration of data from multiple data sources, formats, and integration into an effective consistent data.
- Data mining stage: this stage is the process of data mining analysis, data mining stage is the core of the whole data mining phase. We must first clear mining tasks need to choose what kind of algorithm for mining analysis, clustering analysis and classification analysis; after determining the mining, select the appropriate algorithms for the mining task after pretreatment of data mining analysis, from mining knowledge required by the user.
- Results interpretation and evaluation stage: after the completion of the work in the entire data mining stage, the need to dig out the knowledge is effective for analysis and evaluation, delete the redundant or irrelevant patterns. If the pattern of the excavated is not up to the needs of the user's goals, then you need to back to the previous stage, to data selection, data transformation method changes, sometimes according to the needs of the mining task it is necessary to change a new algorithm for mining association rules.

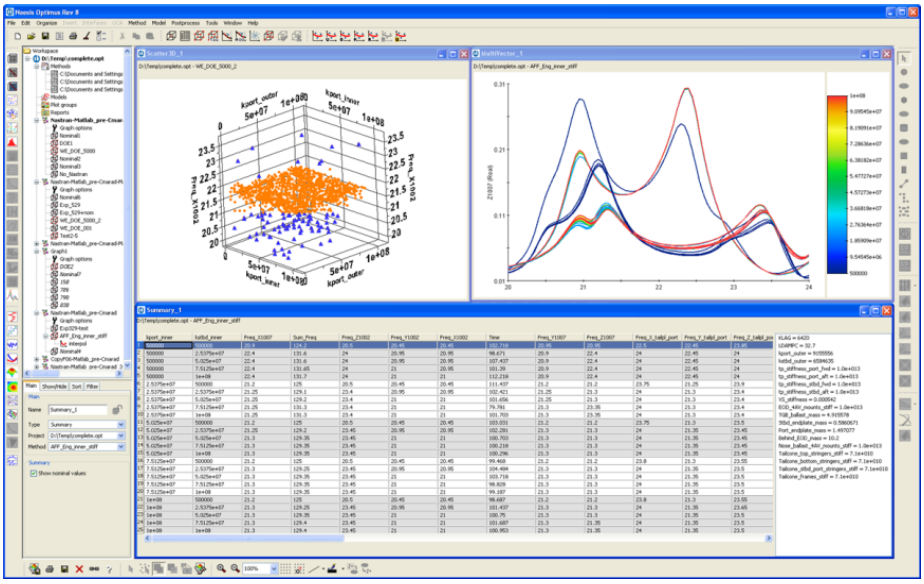


Figure 2. Data Mining Prediction

3. Decision Tree Technology in CET-4 Performance Analysis

3.1. Decision Tree Algorithm

Decision tree is one of the common algorithms, data mining is used for classification, decision tree using top-down recursive way and decision tree internal node attribute value, to determine whether node from the continue to branch down, eventually to the node becomes a leaf node and split ends. Each path from the root node to a leaf node corresponds to a classification rule, corresponding to the decision tree is a set of regular expressions. Decision tree is used in the classification of a tree structure, is a binary tree and binary tree, and data requirements of the input is a set of labeled training data with. Decision trees with multiple internal nodes, each internal node represents a test of the corresponding attribute, decision tree from the root node to each path to a leaf node on behalf of the corresponding attribute of a test result, the decision tree leaf node represents a class or the class distribution, at the top of the decision tree node we call for the root node.

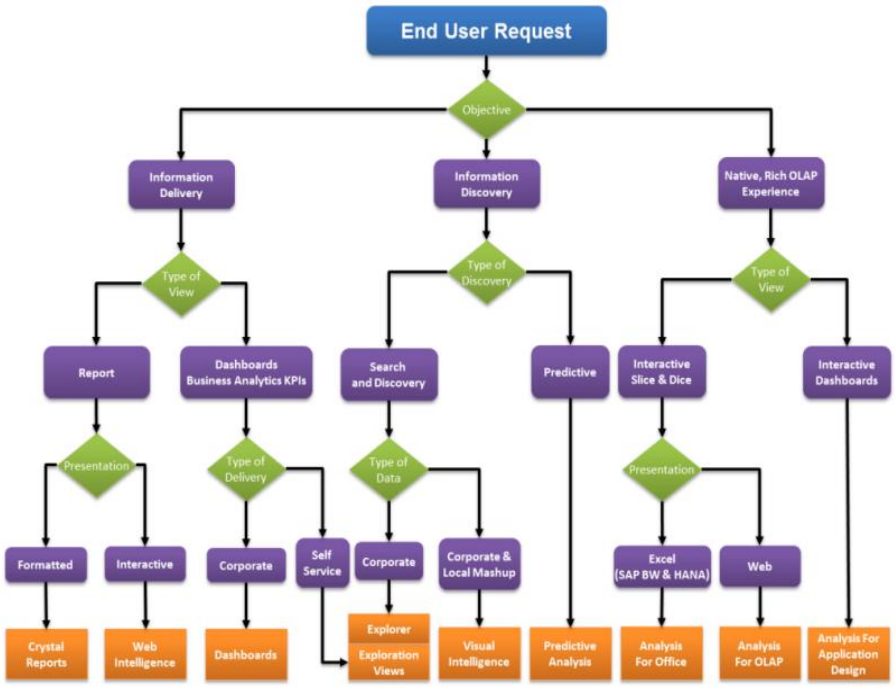


Figure 3. Decision Tree Algorithm

Reduced to the lowest is obtained by computing the information gain of all the attributes, properties of the highest information gain as classification properties, the current node contains sample set divided. This will enable every book produced by concentrated "different types of mixed degree". Algorithm is based on the maximum information gain to select the classification attribute. If the value of the attribute a is divided into T, the training sample set is T2, T1,... , Tm, a total of M subset, then the information gain formula such as

$$Gain(A) = \inf(T) - \sum_{i=1}^{m} \frac{|T_i|}{T} \times \inf(T_i) \tag{1}$$

The training sample set T is classified by the desired information inf (T) calculation method such as the formula:

$$\inf(T) = -\sum_{j=1}^{s} freg(c_j, T) \times \log_2(freq(c_j, T)) \tag{2}$$

Using the gain rate to replace the information gain, the formation of the improved C4.5 algorithm ID3:

$$Gainratio(A) = \frac{Gain(A)}{Split\ \inf(A)} \tag{3}$$

Splitinf (A) represents the resolution of information:

$$Split\ \inf(A) = -\sum_{i=1}^{h} \frac{|T_i|}{|T|} \times \log_2\left(\frac{T_i}{T}\right) \tag{4}$$

3.2. CET-4 Score Analysis

College Students' CET-4 test results not only reflect the students' mastery of English, but also an important indicator to measure the level of College English teaching. English is no longer a simple language, professional skills, but the necessary professional quality of job applicants, all colleges and universities in the effort to try teaching reform, through

the improvement of teaching level, improve the pass rate of CET-4 test. In this paper, the application of decision tree to dig out the main factors that affect students' CET-4 achievements, which is for the improvement of teaching. The research data all from in Jilin University of Finance and economics of the comprehensive educational administration system, the selected data for the study consisted of 5, and a total of 1500 people, data integration is to store the data mining of data source in a new and consistent data storage. In this paper, the data is the basic information of students, student achievement information, students' English CET-4 score information and the basic information of teachers. Using the database technology, the student achievement information table of results calculated public course grade point average, average scores of professional courses, elective courses grade point average, English grade point average; information of CET4 scores screened highest score, finally according to the uniqueness of the student ID, then merge the data into a data table.

According to the principle of the C4.5 algorithm, the decision tree is created by the following steps: calculating the information gain rate of each classified attribute in the training sample set T. Then for each split attribute information gain = rate were compared, find out with the maximum information gain rate of the splitting attribute and set the attribute for the root node of the decision tree, the node has several properties, it is divided into several branches until there is a property with split ends. Repeat steps 1 and 2 for each subset that is split, information gain rate is in Table 1.

Table 1. Information Gain Rate of Each Attribute

attribute	Gain rate
Average information gain rate	0.11045
Information gain rate for admission	0.10680
Average information gain rate of public class	0.05647
Information gain rate of English teachers	0.04329
information gain rate of the four grade examination term	0.03167
Information gain for professional categories	0.02324
Average information gain rate of professional courses	0.02106
Internal and external information gain rate	0.01891

It can be seen from the above results, English grade point average properties in the entire attributes information gain rate maximum, the English column reference properties, effects of factors other than study English achievement factors of CET-4 scores. In the remaining factors admission scores of maximum rate of information gain, therefore to entrance scores as the achievements of the root node, according to the entrance examination of the four kind of different attribute values, the establishment of four branches, then the branching samples for division of property. Then calculate the information gain rate of the other attributes of the branch of the entrance score of 500 to 600 as shown in Table 2.

Table 2. Gain Rate of each Attribute in Branch

Attribute	Gain rate
Average information gain rate of public class	0.06380
Information gain rate of English teachers	0.04541
The information gain rate of the four grade examination term	0.03623
Average information gain rate of professional courses	0.03756
Internal and external information gain rate	0.03244
Information gain for professional categories	0.01215

After the above results can know, entrance examination for maximum rate of 500 to 600 branches in the gender attribute information gain, so choose for the entrance examination for the splitting attribute of the 500 to 600 branches, create a node marked for gender, and according to its attribute values, and then set up a branch. For each subset in each branch, the steps are repeated to determine the individual nodes in the branch, until the subset is empty, and finally the decision tree model is formed.

4. College Students' CET-4 Scores Based on K Mode Algorithm

4.1. K-means Algorithm

K algorithm is the most commonly used algorithm based on partition method in clustering algorithm. The basic idea of the algorithm is to randomly select k objects from a given data set S, which contains n data objects. Each object represents the initial mean or center of a cluster. Then calculate all the data distances to the K centers, each data object into the distance it the centre nearest to where the class, re calculation of data objects of class k mean as each class of the new center, according to a new cluster center re allotment according to the set s, through the iterative cycle, until the criterion function converges, the final clustering result. The square error criterion is usually defined as follows:

$$E = \sum_{i=1}^{k} \sum_{s \in C} \left| S - \overline{S_i} \right|^2 \tag{5}$$

The k-means algorithm in on categorical data clustering analysis may not defined, K mode (K-Modes algorithm offset this disadvantage, it is a variant of the K-means clustering method, using matching dissimilarity to deal with the classification of data objects, cluster mode is used for replacing the cluster mean, with a new dissimilarity measure to deal with object classification, each attribute selection in the category accounted for the largest proportion of attribute values as attributes of the new mode value sequentially updated each new mode until the clustering cost function to achieve optimal, is property of the maximum frequency method updates the cluster mode. The simple matching dissimilarity measure formula is:

$$d\left(x_i, x_j\right) = \sum_{l=1}^{m} \delta\left(x_{il}, x_{j_l}\right) \tag{6}$$

$$\delta\left(x_{il}, x_{j_l}\right) = \begin{cases} 1, & x_{jl} \neq x_{il} \\ 0, & x_{jl} = x_{il} \end{cases} \tag{7}$$

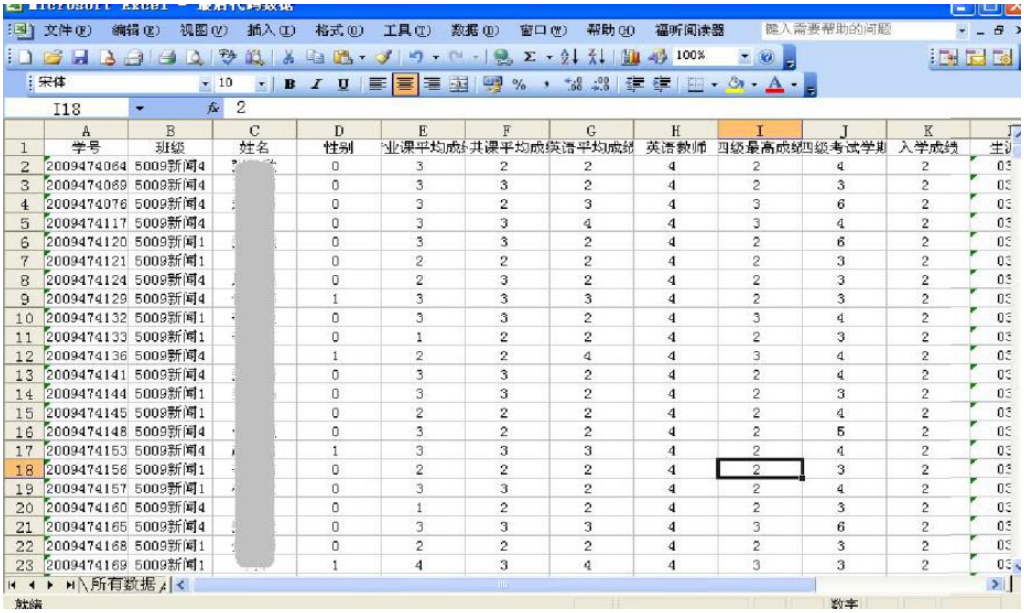
The definition of X is:

$$D(X,Q)=\sum_{i=1}^{n} d(X_i,Q) \tag{8}$$

The criterion for clustering of K-Modes algorithm is to minimize the objective function of the function:

$$\begin{aligned} p(w,z)&=\sum_{l=1}^{k}\sum_{i=1}^{n} w_{i.l}\,d(x_i,z_l) \\ w_{i.l}&\in\{0,1\} \qquad 1\le i\le n \\ \sum_{l=1}^{k} w_{i.l}&=1 \qquad 1\le i\le n \end{aligned} \tag{9}$$

Research of data still use the previous chapter used 1350 data, as shown in Figure 3, then in accordance with CET-4 scores of qualified and unqualified group, qualified table in 815 data objects, not qualified in 535 data object.



	A	B	C	D	E	F	G	H	I	J	K	L
1	学号	班级	姓名	性别	专业课平均成绩	公共课平均成绩	英语平均成绩	英语教师	四级最高成绩	四级考试学期	入学成绩	生源
2	2009474064	5009新闻4		0	3	2	2	4	2	4	2	03
3	2009474069	5009新闻4		0	3	3	2	4	2	3	2	03
4	2009474076	5009新闻4		0	3	2	3	4	3	6	2	03
5	2009474117	5009新闻4		0	3	3	4	4	3	4	2	03
6	2009474120	5009新闻1		0	3	3	2	4	2	6	2	03
7	2009474121	5009新闻1		0	2	2	2	4	2	3	2	03
8	2009474124	5009新闻4		0	2	3	2	4	2	3	2	03
9	2009474129	5009新闻4		1	3	3	3	4	2	3	2	03
10	2009474132	5009新闻1		0	3	3	2	4	3	4	2	03
11	2009474133	5009新闻1		0	1	2	2	4	2	3	2	03
12	2009474136	5009新闻4		1	2	2	4	4	3	4	2	03
13	2009474141	5009新闻4		0	3	3	2	4	2	4	2	03
14	2009474144	5009新闻1		0	3	3	2	4	2	3	2	03
15	2009474145	5009新闻1		0	2	2	2	4	2	4	2	03
16	2009474148	5009新闻4		0	3	2	2	4	2	5	2	03
17	2009474153	5009新闻4		1	3	3	3	4	2	4	2	03
18	2009474156	5009新闻1		0	2	2	2	4	2	3	2	03
19	2009474157	5009新闻1		0	3	3	2	4	2	4	2	03
20	2009474160	5009新闻4		0	1	2	2	4	2	3	2	03
21	2009474165	5009新闻4		0	3	3	3	4	3	6	2	03
22	2009474168	5009新闻1		0	2	2	2	4	2	3	2	03
23	2009474169	5009新闻1		1	4	3	4	4	3	3	2	03

Figure 4. English Achievement Data

4.2. Data Clustering Analysis

To calculate the class center, that is in each class, for each attribute selection in the accounts for the largest proportion of attribute value as a mode of the class attribute value. With the "gender" attribute, for example: in cluster 1 female 429, men and 142, percentage of women accounted for 75.1%, percentage of men accounted for 24.9%, so 1 cluster mode in "gender" attribute set to "female". According to the characteristics of this group of students, the value of K will not be too large, so this paper in the K value of the minimum value is 2, the maximum value of 10. According to the calculation method of K for 2, the clustering results of K for 3-10 are obtained in turn. Through the analysis of the final clustering result data statistics and comparison, combined with years of student management work experience judgment, when the initial K value for 5 effect is ideal, the qualified table of cluster analysis, the program after 11 iterations to generate a clustering results: 20 objects on the cluster 1, cluster 2 116 objects, cluster 3 135 object, Cluster 4 169 object, cluster 5 375 object. The clustering results are shown in Table 3.

Table 3. Results of k=5 Clustering

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of objects	20	116	135	169	375

Cluster 1 has 20 instances, accounting for 2.5% of the total. The main achievement in English; admission scores mainly between 400-500; girls accounted for 65%; Public Course good majority. Cluster 2 has 116 instances, accounting for 14% of the total. English score 53% as medium; and 83.6% of the students admission scores between 400-500; the proportion of men and women 50-50; the proportion of each semester exam rather, each layer has a segment scores; professional Division based on information. Cluster 3 has 135 instances, accounting for 16.6% of the total. English results excellent, good; 80% of the students admission scores between 400-500; female proportion was 70%; Public Course, excellent, good, each accounting for about 50%; the third term by the proportion of 55% other semester exams by those roughly; scores at all levels have, gifted a share of 44%.There are 169 instances in cluster 4, 20.7% of the total. English score of 58% as good; 78% of the students admission scores between 400-500; 84% were women; Public Course relies mainly on good, 53% of students in the third through the semester; Scores 50%.There are 375 instances in cluster 5,46 per cent of the total. English results good, excellent; 80% of the students admission scores between 400-500; 271 girls who share 72%; Public Course in benign; Scores at all levels have, excellent, good, each about 25%; 56% of the students in the third semester by four exams.

5. Conclusions

With the information of the management of colleges and universities, some new educational management system, performance management system, school management system have been put into use, to the day-to-day management of the school work has brought great convenience, improve the management efficiency and management level. After many years of accumulation, these management system stores a lot of valuable information resources, more and more universities and researchers into use the electronic data to discover between the grades of students and the teaching of some rules and knowledge, provide scientific basis and teaching decision support for the managers.

Students' achievement is not only an important sign to measure students' mastery of knowledge, but also an important basis to measure the level of education and teaching in a college. With the expansion of the scale of running school, the increase in the number of students in school, to the education and teaching management has brought a lot of pressure. At present, students to participate in the examination has accumulated a lot of electronic performance data, such as the school test scores, CET-4 test scores. These data are stored in the database, providing a simple query, modify and statistics functions, whether there is a link between the data cannot be determined. Through this system can on student achievement of mining analysis, the learning situation of students have more comprehensive understanding and the understanding, further analysis of students' grasp of knowledge, to identify factors that influence student achievement, and guide students to check leakage fill a vacancy, enhance the efficiency of learning; promote the teachers to improve the problems existing in the teaching, improve teaching quality; managers to provide decision-making basis.

Acknowledgments

It is funded by the eighth batch of Chinese Foreign Language Education Fund Project “Research of College English Curriculum and Teaching Method Based on the Corrective Feedback Mechanism”, (approved number: ZGWYJYJJ2016B06).

References

[1] Ergun,G.,Wu,M. The influence of internationalisation of higher education: A China's study,Procedia - Social and Behavioral Sciences, Vol. 2, (2010), pp .5675-5681.
[2] Fornell, C., Larcker, D. Structural equation models with unobservable variables and measurement error:Algebra and statistics. Journal of marketing research, Vol. 18, (1981) , pp .382-389.
[3] González,M., González, L. The co-creation as a strategy to address IT governance in an organization.RISTI-RevistaIbérica de Sistemas e Tecnologias de Informação, Vol. 14, (2015) , pp .1-15.
[4] Hu,M., Xu,S. Research of Multimedia Teaching on Principles of Management.IERI Procedia, Vol. 2, (2012) , pp . 666-670.
[5] Ihfasuziella, I.,Wan,Z. Space Management: A Study on Space Usage Level in Higher Education Institutions.Procedia - Social and Behavioral Sciences, Vol. 47, (2012) , pp .1880-1887.
[6] Marjan, L. Knowledge management in higher education. Procedia Computer Science, Vol. 3, (2011) , pp .544-549.
[7] Oana,M. The implementation of quality management in higher education.Procedia - Social and Behavioral Sciences, Vol. 5, (2011) , pp .1046-1050.
[8] Pradit,S. The Knowledge Management in Higher Education in Chiang Mai: A Comparative Review. Procedia - Social and Behavioral Sciences, Vol. 69, (2012) , pp .399-403.
[9] Wei, C.,Tao, Y. Application of Multimedia–Aided Project–Teaching Mode in Cultural Education.IERI Procedia, Vol. 2, (2012) , pp .538-542.

Author

Xiao Yanjiao, 1981.06 Luotian, Hubei, China. Current position: the lecturer of College of Foreign Languages, Southwest Petroleum University, Chengdu, China. Scientific interest: Her research interest fields include English education, English and American literature and culture study. Publications: more than 10 papers published. Experience: She has 10 years of teaching experience and has completed five scientific research projects.

