Research on the Application of Data Mining in the Analysis of College English Teaching Quality

Xiaohua Guo1,*

1Foreign Language Department, Ganzhou Teachers College, China

*Corresponding author e-mail: guoxiaohua@ganzhsz.cn

Abstract. This article uses the Apriori algorithm in the analysis of college English teaching quality, and uses this algorithm to analyze the learning situation of 1133 students. From the various elements including the college English teaching period, prerequisites and learning environment, the research shows that the main factors that have the greatest impact on learning are learning motivation, teaching methods and teaching modes. Finally, the experimental analysis shows that the Apriori algorithm used in this article can quantify the impact of different teaching factors on students' learning quality, and also provide a basis for subsequent research on the quality of college English teaching.

Keywords: English Teaching, Quality Improvement, Data Mining, Efficiency

1. Introduction

Teaching quality evaluation is the process of systematically collecting information according to the requirements of teaching goals and teaching principles, and giving value judgments to teaching activities and teaching results in the teaching process[1-2]. At present, the teaching evaluation methods of colleges and universities are not all the same, which can be roughly divided into two categories: one is a single qualitative evaluation. Qualitative evaluation is a traditional evaluation method, which mainly focuses on qualitative analysis. This method is too flexible and too rough to reflect the actual situation of teaching objectively, accurately and in-depth, and is far from the requirements of modern teaching management for teaching evaluation[3]. Quantitative evaluation mainly reflects the status of teaching quality through quantitative analysis. However, the current quantitative evaluation methods have the following problems: the evaluation content is too simple; the evaluation subject is too single; the evaluation method is single[4-5]. With the in-depth development of internal teaching quality evaluation activities in my country’s colleges and universities, how to express, analyze, interpret, publish and use the results of teaching evaluation can arouse the enthusiasm of teachers and students and make evaluation play a greater role. New research topics[6]. Data mining technology is incomparable to other technical methods in mining the laws implicit in own data and solving specific problems. It can discover many potentially valuable associated information in the evaluation information, help the teaching department to make decisions, and provide valuable references for teachers.

Based on this, this article believes that objectively and accurately discovering important factors
affecting the quality of college English teaching is the premise and basis for constructing a reasonable teaching quality evaluation system and improving teaching in a targeted manner. This paper proposes an analysis method of teaching quality factors using data mining technology. Through the analysis of teaching link data, we try to find quality influencing factors hidden in the teaching process, conditions and environment, and provide a new way for teaching quality analysis and improvement. Thoughts and methods. Data Mining (DM) is a data retrieval method that uses statistical algorithms to find certain laws and related knowledge in data. It is an effective way to discover knowledge from massive data and is widely used in business decision-making and management. Practice. This paper uses data mining methods to analyze teaching quality factors and discover knowledge in teaching-oriented data warehouses.

2. Experimental Design

2.1. Subject and purpose
From April to June 2007, 1132 non-English majors in the first year of a university were selected as the survey subjects, and the making notable progress (MNP) of individual students was used as the objective evaluation index. The Apriori (CHi -squared Automatic Interaction Detector) algorithm analyzes the factors affecting MNP.

2.2. Experimental method
This article uses the Apriori algorithm to analyze the factors that affect MNP. Apriori is an analysis method based on target optimization, with target selection, variable screening, and clustering functions. It is suitable for classification and analysis of ranking data. It optimally divides the sample according to the given response variable and the filtered explanatory variable, and automatically judges and groups the multivariate contingency table according to the significance of the chi-square test. It usually has a good effect on the automatic classification of discrete data sets. In view of the limitations of the Apriori algorithm, the time efficiency of the algorithm and the storage space requirements, the Apriori algorithm is improved in this article.

2.3. Use the screening method to further reduce the number of candidates in the candidate project set
In the Apriori algorithm, after Ck-1 is generated, it is compared with the support. The itemsets smaller than the support will be pruned to generate Lk-1, and Lk-1 is connected with Lk-1 to produce Ck. The improved algorithm is to further filter Lk-1 before the candidate item set Ck is generated, count the number of occurrences of all items in Lk-1, and delete the item set containing items with less than k-1 occurrence in Lk-1 to reduce The number of k-1 itemsets participating in the connection, so as to reduce the number of candidates in Ck.

2.4. Using support transaction intersection method to reduce the number of scans of the transaction database
First, scan the source database to record the support affairs of each item. When seeking the support of a certain candidate item set, first intersect the support affairs sets of all items in the item set to obtain the support of the candidate item set. Transaction set, and then find its support. While scanning the database, record the support affairs of each project, and determine the support transaction set of the candidate k-item set through the support transaction set of each project in the candidate k-item set, which will avoid repeated scanning of the database and greatly save the system. Overhead.

2.5. Improved algorithm description
The candidate 2-item sets are obtained by concatenating the items in the frequent 1-item sets in pairs, and then the support transaction sets of each set are obtained by intersecting the support transaction sets of the two items in the 2-item sets. Delete the itemsets whose number of supported transactions is less than the minimum supported transaction, and get frequent 2-itemsets.
Before the candidate set is generated, the items in the frequent k-items set are deleted by filtering, and the resulting set is called the valuable frequent k-item set, and then according to the connection step in the Apriori algorithm, the items in the valuable frequent k-item set Items are processed to obtain the candidate (k+1)-item set, and the support transaction set of each item is obtained by intersecting the support transaction set of each item in the candidate (k+1)-item set, and the support is deleted. The number of transactions is less than the item set of the smallest supported transaction, resulting in frequent (k+1)-items set.

4) Repeat the operation 3) until the frequent k-items set is empty.

That is, the optimal sample \( (1,1,\cdots,1)^T \), n+1 data sequences form the following matrix:

\[
\begin{bmatrix}
X_0'(1) & \cdots & X_s'(1) \\
\vdots & & \vdots \\
X_0'(m) & \cdots & X_s'(m) \\
\end{bmatrix}_{m=n+1} \tag{1}
\]

\[
\begin{bmatrix}
X_0(1) & \cdots & X_s(1) \\
\vdots & & \vdots \\
X_0(m) & \cdots & X_s(m) \\
\end{bmatrix}_{m=n+1} \tag{2}
\]

2.6. Sample characteristics

In order to systematically observe the teaching factors that affect students' learning quality, we designed 13 attributes in the questionnaire that may affect teaching quality, such as "learning purpose", "teaching means", "teaching mode", and "teacher type". Some of these attributes reflect the subjective factors in the learning process of students, such as "learning purpose", "overall evaluation of teachers", "entry English scores", etc.; while others reflect objective factors such as teaching conditions and environment. Such as "teaching mode", "teaching means" and "teacher qualifications".

3. Experimental results and analysis

3.1. The importance of the factors affecting the quality of learning

In this paper, the experimental results are expressed in the shape of a classification decision tree. The decision tree has 52 nodes at the 95% confidence level, and this paper intercepts 15 important nodes.

This article uses the Percentage of MNP (PMNP) to express the quality of learning. Through this decision tree, the importance of the various factors affecting PMNP can be found. Since the Apriori algorithm is to automatically judge and group the multivariate contingency table according to the significance of the chi-square test, the classification factor (ie, influencing factor) closer to the root of the decision tree is more important.

Among the 1132 student samples, 208 have made significant progress in their test scores compared with the previous test. The total PMNP of all samples is 18.4%. The most important factor affecting PMNP is the "learning purpose"—for example, the PMNP of students whose learning purpose is "interest" is nearly double that of students who "pass the test." According to the classification order of the decision tree, the second most important factor that affects PMNP is "teaching method", again is "teaching mode", and other factors are second.

3.2. Analysis of the influence of various factors

Each node in the decision tree is a specific classification group obtained through Apriori operation. The result data reflects the influence of various factors on PMNP, which can be described in the form of production rules, namely "if...then...".

For example, the rule reflected by node 14 can be described as:

Through the data of each node of the decision tree, it reflects the influence of various factors on
PMNP, which is reflected in two aspects:

1) Each node reflects the PMNP results under various classification methods, allowing analysts to intuitively understand the teaching quality under different factors and environments of English teaching (such as teaching modes, teaching methods, and teacher education).

2) The statistical data indicators of each node (as shown in Table 1) can further and quantitatively reflect the actual effect of each factor. For example, of the 44 samples included in node 8 in Table 2, 13 have reached MNP, their PMNP is 29.5%, and the PMNP rate is 1.61. This shows that the PMNP of the classification group represented by node 8 (that is, the learning motivation is "finding a job or going abroad" and the teaching method is "rich multimedia") is 61% higher than the average PMNP of the sample complete set.

Table 1. Statistical data indicators of some nodes.

| Node number | A: The number of samples contained in the node | B: A accounts for the total, Percentage of the sample (%) | C: The number of samples with significant improvement in academic performance | D: Percentage of C in the total sample (%) | PMNP(%) | E: PMNP magnification E |
|-------------|-----------------------------------------------|--------------------------------------------------------|-----------------------------------------------------------------|----------------------------------------|---------|-------------------------|
| 8           | 45                                             | 4                                                      | 14                                                               | 6.4                                     | 29.6    | 1.61                    |
| 6           | 18                                             | 1.6                                                   | 6                                                                | 2.5                                     | 29.5    | 1.6                     |
| 5           | 291                                            | 25.7                                                  | 72                                                               | 34.2                                    | 24.6    | 1.33                    |
| 1           | 595                                            | 52.6                                                  | 133                                                              | 63.6                                    | 22.3    | 1.21                    |
| 4           | 305                                            | 27                                                    | 62                                                               | 29.4                                    | 20.2    | 1.09                    |
| 3           | 163                                            | 14.4                                                  | 32                                                               | 15                                      | 19.2    | 1.04                    |
| 21          | 89                                             | 7.9                                                   | 17                                                               | 7.8                                     | 18.3    | 0.99                    |
| 18          | 13                                             | 1.2                                                   | 3                                                                | 1.1                                     | 16.8    | 0.91                    |
| 9           | 119                                            | 10.5                                                  | 19                                                               | 8.8                                     | 15.4    | 0.83                    |
| 14          | 152                                            | 13.4                                                  | 23                                                               | 10.7                                    | 14.7    | 0.79                    |
| 13          | 217                                            | 19.2                                                  | 24                                                               | 11.2                                    | 10.7    | 0.58                    |

A is the number of samples contained in node N; B is the number of samples in node N A/total number of samples, expressed as a percentage; C is the number of samples with significant improvement in learning performance in node N; D is C/remarkable improvement in learning performance The total number of samples, expressed as a percentage; E represents the multiple of the PMNP of node N relative to the PMNP of all samples (ie 208/1132).

3.3. Targeted teaching quality improvement strategies

Through PMNP decision tree analysis, it can help decision makers to propose targeted teaching improvement measures.

For example, node 13 indicates that even if the student’s learning motivation is “interest” (this is the group with the highest average PMNP among all learning motivations), but in the “Changpu A” teaching mode and the teaching method is “multimedia without or insufficient”, Its PMNP is only 0.58, which is in sharp contrast with node 18. Under the specific teaching mode of "Changpu A", the influence of multimedia teaching methods on academic performance is not More significant. The explanation for this phenomenon is that the basic English of Changpu A students is generally not as good as that of "Happy English" and "High Starting Point", especially in the basic aspects of listening and speaking. Therefore, this group of students is more familiar with computers such as multimedia teaching. The dependence on auxiliary teaching tools is greater. For such students, the use of multimedia teaching tools should be particularly strengthened, and high-quality teaching courseware covering all aspects of "listening, speaking, reading, and writing" should be developed.
4. Conclusion
This paper proposes an analysis method of college English teaching quality based on data mining algorithms, using the Apriori algorithm according to the rapid progress rate of students' performance as an analysis index of teaching quality. The algorithm uses data mining algorithms to find out the factors that are contained, extremely hidden, but have important research value from including the college English teaching period, preconditions and learning environment. Experimental research shows that the method proposed in this article is beneficial to the quality of college English teaching and provides effective strategies to compare the pros and cons of different teaching measures or the applicability of different student groups. The analysis of college English teaching quality provides a basis.

References
[1] Chunping, N., & Xin, G. (2014). Analysis of the application of a happy learning method in college english classroom teaching. International Journal of Technology Management(4), 11-13.
[2] Zhang, D., & Jiang, K. (2012). Application of data mining techniques in the analysis of fire incidents. Procedia Engineering, 43, 250-256.
[3] Buenano-Fernandez, D., Villegas-Ch, W., & Lujan-Mora, S. (2019). The use of tools of data mining to decision making in engineering education-a systematic mapping study. Computer applications in engineering education, 27(3), 744-758.
[4] Danping, Z., & Jin, D. (2011). The data mining of the human resources data warehouse in university based on association rule. Journal of Computers, 6(1), 139-146.
[5] Meyer, B., & Linden, G. (2011). In support of open reviews; better teaching through large-scale data mining. Communications of the ACM, 54(11), 12-13.
[6] Xu, G., & Li, L.. (2013). Social media mining and social network analysis: emerging research. Online Information Review, 38(1), 157-158.