Research on the Storage and Analysis of Teaching State Data of the University

Min LI¹,* and Xiang-jun DU²

¹Teaching Affairs Office, Qingdao University, Qingdao, Shandong, China
²College of Computer Science & Technology, Qingdao University, Qingdao, Shandong, China

*Corresponding author

Keywords: Teaching state data, Data mart, Data mining, Complex network.

Abstract. Teaching state data of the university has many characteristics like big data, such as volume, variety, value, velocity and so on. It is difficult to cope with data integration, multi-dimension, multi-granularity, systemic analysis and mining. This paper presents the storage model and analysis framework of teaching state data based on data warehouse, data mining and complex network theory, and collects teaching state data accumulated by Qingdao University. Further, a multi-level data mart is proposed to store the data, and OLAP, association rule mining, clustering analysis as well as complex network modelling tools are developed to analyze these data. The analysis of test result data and teaching quality evaluation index data show that the model and the framework are helpful for knowledge discovery of teaching state data.

Introduction

With the rapid development of information technology, information system is more and more applied to the daily management of various fields, and these information systems produce a large number of data. The teaching state data of colleges and universities are produced by various teaching management systems, which reflect many aspects of the teaching process in colleges and universities. It can be used to measure the quality of teaching and management, and play a pivotal role in the undergraduate teaching evaluation [1]. In view of this, the Ministry of Education of China has developed a national data platform for quality inspection of higher education, constructed a data collection index system for teaching state in colleges and universities, and collected the basic teaching state data of colleges and universities every year, and collated and measured the data to produce the analysis report of basic teaching state data for each specific college or university as the basis for teaching evaluation. These analysis reports are based on the data to describe and explain the teaching work of colleges and universities, and provide a strong support for the comprehensive and systematic understanding of the basic situation of the school, and improve the validity and reliability of the teaching evaluation.

On the other hand, the teaching state data reflects the regular work of many teaching departments in colleges and universities, is usually stored in many isolated and heterogeneous systems. The correlation among the data produced from different systems also makes the relationship between data complex and the scale of data growing rapidly. These massive, complex teaching state data indicate that teaching management in colleges and universities has entered the era of big data [2,3,4]. However, the traditional database-based data storage and statistical analysis methods are often limited in dealing with such massive and complex teaching state data, because it is difficult to effectively cope with the massive and complex data set with them. It makes that the effective analytical tools for teaching state data in multi-dimensional, deep-seated, systematic mining are still insufficient, the valuable information hidden in teaching state data has not been fully revealed and the role of teaching state data in teaching evaluation and teaching reform can be further strengthened. Therefore, to effectively integrate the massive, heterogeneous, scattered and rapidly growing data accumulated in the daily teaching process of colleges and universities, and then analyze and mine the data to obtain the teaching rules hidden in the data, and finally improve the quality of teaching evaluation and teaching reform, has become an urgent problem to be solved.
Based on the data accumulated by Qingdao University for many years, we study the storage model and integration method of teaching state data based on data warehouse, and present a comprehensive analysis framework based on the OLAP (On-Line Analysis and Processing) technology, data mining and complex network modelling for the analysis of teaching state data, and develop many analyzing tools. After that, we apply these tools to analyze the teaching state data of Qingdao University and obtain some valuable information, which are helpful for extracting teaching quality evaluation index and improving teaching evaluation.

The Storage of Teaching State Data

As an important storage form of large data, the data warehouse is subject-oriented and application-oriented. These features make it necessary to construct the data warehouse and integrate the data from the heterogeneous data source into the data warehouse after obtaining the theme focused by the users, and analysis and mining can be carried out only after the storage has been constructed and the data has been integrated. Therefore, the storage of teaching state data is the basis of analysis and mining as well as other applications.

Multidimensional Storage of Teaching State Data

Because most teaching information management systems of colleges and universities mainly use the relational database to store the data [5], constructing the data warehouse based on the relational database contributes to the extraction, transforming and loading of the data. However, the join operation between multiple tables is necessary for accessing relational database, and its efficiency is low because of the characteristics of "high dimension" and "long dimension" of teaching state data. So it is difficult to realize the quick query performance of data analysis and decision making in teaching management and evaluation. Therefore, we combine the multidimensional storage and relational storage in the construction of teaching state data warehouse, and develop the data warehouse based on the relational database to store the detailed data with the low query frequency, and develop the data warehouse based on multidimensional model to store the smaller scale of the aggregated data with the high query frequency. Figure 1 is an illustration for the multidimensional cube of the test results. In the cube, three dimensions, like the college, the grade and the curriculum, and a measure index for the test results, are extracted and constructed. In addition, a small cube with bold dotted borders in Figure 1 means the average score of the high mathematic curriculum of Automation College in 2009. As shown in Figure 1, with this cube, we can easily analyze and aggregate the rest results from the three dimensions.

Multi-level Data Mart Storage of Teaching State Data

For the teaching state data, there is a certain hierarchy in some thematic data. These hierarchical relationships can be used to optimize the storage and contribute to the improvement of query
performance. For example, by querying the detailed data about teachers in the teaching state data, one can get all the data about any one teacher. Moreover, one can also obtain the aggregated data about all teachers belonging to a certain college through a coarse-grained query according to the institution dimension attached to the teacher. The difference in the granularity between the thematic data makes there are obvious hierarchical relationships in the teaching state data. In order to construct the storage of the data with such hierarchical relationship and further to use the hierarchical relationship between the thematic data to improve the query performance for the teaching state data, we propose a multi-level data mart storage model. In this model, we first extract the same attributes of all the data marts to be analyzed and model these attributes to the dimensions, and then build the level and hierarchy for these dimensions according to the hierarchical relationship among these data marts in the dimensions that have been built. In other words, the hierarchy of a dimension is composed by some corresponding levels attached to this dimension. All cubes are organized into a tree structure according to the hierarchy of a dimension, and a cube is a data mart with a specific level in the hierarchy. Furthermore, we build and materialize all data marts for teaching state data to be analyzed, and apply the multi-level data mart model to build a hierarchy for all data marts that have been built, in order to improve the query performance through a support from a dynamically optimized access policies.

Through the analysis on the teaching state data of colleges and universities and the actual teaching management of Qingdao University, we construct a three-level data mart storage. In this storage, all data marts are mapped into three levels according to the scale of the department: University → College (Department) → Profession, and three levels in the time dimension: School year → Semester → Month.

The Analysis of Teaching State Data

The traditional teaching management system mainly relies on the relational database or the flat file data [5], and its basic operations include the ‘insert’, the ‘delete’, the ‘update’ and the ‘select’. With the continuous increasing information management systems adopted in colleges and universities, massive, heterogeneous, multi-dimensional and multi-granularity teaching data has aroused people's attention, and the OLAP and many data mining methods have also gradually applied in the analysis of teaching data. Studies have shown that OLAP contributes to the aggregation and multi-dimensional online analysis of massive data, and data mining is very effective in revealing the rich information hidden in massive data, especially for teaching management of colleges and universities [6]. However, teaching state data of colleges and universities also show strong complexity, which mainly relies on the complex interaction among massive entities in the data. The information hidden in the complex relationship between the entities is of great significance for teaching management and teaching evaluation.

Complex network is one of the main methods to study the complex system. By modeling the individuals of a system into nodes, and the relationships between individuals into edges, complex network effectively reflects the structure of a complex system. Furthermore, the statistical analysis and evolutionary analysis of complex network can be used to reveal the function of the system. At present, a few studies have tried to apply the complex network model to the analysis of teaching data in colleges and universities [7].

Based on the above analysis, we consider that OLAP, data mining and complex network play different role in the analysis of teaching data. OLAP mainly provides multi-dimensional and multi-granularity analysis of numerical measures, and data mining helps people to discover hidden information in massive data, while complex network focuses on the analysis of complex relationships between entities in the data. Although three methods contribute to the analysis of teaching data, it is very scarce to combine OLAP, data mining and complex network into a framework to systematically analyze teaching state data.

Therefore, based on the multidimensional storage and multi-level data marts described in section 2, we develop a series of multidimensional online analysis tools, association analysis tools, cluster analysis tools and complex network analysis tools for teaching state data. Then we integrate these tools and develop a teaching state data analysis system of Qingdao University. Using this system, we analyze and mine teaching state data accumulated in Qingdao University for many years, and obtain some valuable information.
Multidimensional Analysis of Teaching State Data

In the teaching management of colleges and universities, the analysis of teaching state data involves the daily work of various departments. The teaching managers of different roles such as universities and colleges usually have different requirements. Therefore, a flexible, diverse, user-friendly OLAP tool is needed.

We develop a Browser/Server-based OLAP tool, which is embedded into the flexible Web page. To improve the performance of OLAP, we develop a server-side middleware, which aims to support for fast multi-dimensional access. In order to improve the transmission efficiency of query results, we compress all the result data about multidimensional query and OLAP analysis in the server and then transmit these compressed data to the client. Based on that, we develop many visual operations to implement the rolling, drilling, slicing and other operations of traditional OLAP by some simple mouse operations, including the click, the drag and so on, and also provide a variety of graphics and reports to the user. All of these help the teaching manager to analyze teaching state data.

Figure 2. An illustration to the application of OLAP tool in the analysis of test results.

Figure 2 is an illustration to the application of our OLAP tool. Figure 2 a) intuitively reflects a significant difference among the average scores of all students from different provinces in a same course. From that, we can found that the average score of students from Shandong province where Qingdao University locates is far lower than that of students from other provinces. Figure 2 b) reflects the distribution of the average score of all students in four grades from 2007 to 2010. We found that the average score decreases gradually with the increase of grade.

Knowledge Discovery of Teaching State Data

In recent years, the data mining technology has aroused the attention of teaching managers and has become an important means of analyzing and processing teaching state data [8]. In the teaching state data, there is a rich correlation between data entities, and the mining of association rules based on multidimensional cube is an important problem. Li et al. [9] has applied the Apriori algorithm to the frequent predicate search of the data cube, and then extracted the association rules with the frequent predicate. In [10], it is pointed out that each dimension in the multidimensional cube has a certain hierarchy, and the sparseness of the multidimensional data makes it difficult to find strong association rules between the lower levels. Therefore, it is possible to use the drilling, the rolling and other operations to adjust the level to facilitate the discovery of strong association rules. In [11], it is proposed that the efficiency of multi-dimensional association rule mining can be improved by "secondary pruning" method. Based on the above research results, for the multi-dimensional association information enrichment in the teaching state data, we also optimize the Apriori algorithm and apply it to association rule mining of the teaching state data, in order to implement the multidimensional correlation analysis of the thematic data mart. Taking the association rule mining of the data about the teaching management evaluation index as an example, we use the above algorithm.
to obtain 312 association rules and measure the correlation between the 44 indexes by setting the support degree to 0.2 and the confidence degree to 0.8.

Clustering analysis is one of the important methods in data mining. In the face of the massive historical data accumulated in the teaching management, in order to obtain the potential valuable information, it is often necessary to identify the categories of data entities. The clustering analysis can divide the data into different clusters according to the similarity of the data, so that the data entities to be grouped into a cluster are as similar as possible, and the differences among the data entities to be grouped into different clusters are as large as possible. In our system, the clustering analysis tools are developed based on two algorithms, including K-means and EM. The former is easy to implement and can be scalable and efficient to handle teaching state data. However, due to the diversity of the data integrated into our system, although the data has been preprocessed, there may still be a large number of "noisy" and incomplete data, but clustering analysis based on K-means is not suitable for such data. Therefore, we implement another clustering analysis tool based on the EM algorithm. Because the EM algorithm is based on the probability of the membership of the data entity and the cluster, it can be used to mine the incomplete or "noisy" data. By combining the two clustering tools together, teaching managers can group teaching state data into different clusters to discover their potential category patterns in order to support teaching decision making. Figure 3 a) is an example illustration to apply our data mining tools for the clustering results of the scores of 111 students in international trade profession in 2006. By setting the number of clusters to 3, the student score is divided into three levels. By analyzing the centerline of the cluster line (shown in Figure 3 b)), cluster 2 performs better, cluster 0 general, cluster 1 poor. According to the results of the above analysis, the score of all students can be divided into three levels, and it may be helpful for student assessment and teaching quality evaluation.

![Figure 3. An illustration to the application of cluster analysis tools in the score theme.]

**Complex Network Modelling and Analysis of Teaching State Data**

Complex network is a model that describes the composition and structure of complex systems. In the model, the system components are usually modeled as nodes and the relationships between system components is modeled as edges. Studying the statistical characteristics and evolution of the model is helpful to reveal the function of system. In the complex data set, there are often a large number of data entities. These data entities are diverse and the relationships among the entities are often implicit. Therefore, it is an effective method to analyze complex data sets by using complex networks.

Based on the association rules obtained by mining teaching management evaluation index data, we build a correlation network of teaching management evaluation index. In this network, indexes in the association rules are modeled as nodes, the implied relationships between the evaluation indexes are modeled as edges and the confidences of the rules are modeled as the weights of corresponding edges. Then we calculate the degree distribution and the average path length of this network. By calculation, we find that the average path length is 1.55, and it shows that these indexes are closely correlated. In all indexes, the out degree of the node corresponding to the index "The ratio of the number of professors and associate professors charging the theoretical teaching for undergraduate and junior college student on the number of all professors and associate professors" is 15, and it is the maximum of the out degrees of all nodes. This result shows that the index "The ratio of the number of professors and associate professors charging the theoretical teaching for undergraduate and junior college student on the number of all professors and associate professors" has the greatest impact on other indexes, and it should be given the greatest weight in the construction of the index system.
Conclusions and Future

The storage and analysis of teaching state data is a key problem in the teaching quality evaluation and teaching evaluation in colleges and universities. Based on data warehouse, data mining and complex network theory, we design a multi-dimensional model and multi-level data mart model for the storage of teaching state data. Furthermore, we develop an OLAP tool, two association analysis tools, two cluster analysis tools and a complex network analysis tool, and integrate these tools into the teaching state data analysis system of Qingdao University. We apply the system to the daily teaching management and teaching quality evaluation of Qingdao University, and mine some valuable information which contributes to teaching management and the analysis of teaching quality evaluation index.

However, the integration and analysis methods of the teaching state data we have developed can be still further studied in terms of the integration, the efficiency and so on. We will try to fix these problems in the further study.

Acknowledgement

This research was financially supported by the undergraduate university teaching reform research project of Shandong Province (NO. 2015Z044) and the university science and technology plan project of Shandong province (NO. J15LN40).

References


