

Available online at <http://www.mecspress.net/ijeme>

A MDA-based campus data analysis and visualization framework

Jiangning Xie, Xueqing Li, Lei Wang, Yuzhen Niu

School of Computer Science and Technology Shandong University Jinan, 250101, China

Abstract

As the rapid development of high education informatization, it will be a new important research topic to conduct the data mining and visual analytics to the existing data in the separated information systems. The existing campus information system is the integration of some business systems, so the system has some problems, such as it stores the data separately and it has poor ability to analyze data. In order to solve these problems, we propose Model Driven Architecture (MDA) based campus data analysis and visualization framework. The framework is composed of multi-dimension data modeling, data extraction, visualization-based data exhibition, and some other modules. Data extraction solves the problem caused by separately stored data and heterogeneous data. Multi-dimension data modeling analysis and visualization enhances the analysis ability of existing system. Based on the idea of MDA modeling analysis, we provide a rapid develop platform of campus business for both business analyzers and developers.

Index Terms: Data analysis; MDA; Data visualization; Data mining; Scatter Charts

© 2012 Published by MECS Publisher. Selection and/or peer review under responsibility of the International Conference on E-Business System and Education Technology

1. Introduction

As the rapid development of computer techniques and network techniques, network-based school business information management systems, such as enrollment management, teaching affairs management, graduate management, financial management, are extensively used in colleges and universities. Furthermore, the scale and functions of these systems are increasing. As the increasing of enrollment in colleges and universities, information management is showing more and more advantages. Many colleges and universities are using student management information system whose basic functions include enrollment management, training management, course and score management, graduate management, employment management, and so on. In this way, standardized, information-based, and network based students management is achieved. At the same time, a large amount of information and data are accumulated during the daily work. And these systems can storage, backup, query, and statistic the data. However, there are still some questions listed in the following:

1. During the process of management information system, all of the management phases accumulate large amounts of data, but different data is stored in different database, so the data is not well utilized.

* Corresponding author.

E-mail address: xjn@sdu.edu.cn, xqli@sdu.edu.cn, wanglpqqq@gmail.com, yuzhenniu@gmail.com

2. The management system has the basic management functions of data, but its ability to analyze data is weak, it cannot analyze and statistic data from multiple perspectives, it cannot mining the mass data and got useful information, and it cannot provide enough information to colleges and universities business managers for them to make the right decisions.
3. The statistic and analyze functions of the management system are simple and the analysis results are exhibited as reports and tables which are not intuitive. The existing analyze functions are limited in counting the number of people, courses, and summarizing the scores. Such analysis usually cannot neither show potential problems, nor exhibit the inter relationship within data.

So the key points of campus data analysis and visualization are analyze students management data from multiple perspectives and using statistical analysis and data mining methods [1], provide accurate, intuitive, and useful information to support all kinds of department manages to make correct decisions. In paper [2] and [3], the authors present campus information management methods that combine data mining and data analysis. But each method only aims to solve certain business problem. These methods cannot analyze the whole business data, and don't have any visual analytics function, so cannot provide any intuitive visualization.

In order to solve the problems of data analysis and visualization, based on data mining techniques, data visualization techniques, and MDA [4], we propose a campus data analysis and visualization framework. Base on MDA, the framework gives the models of multi-dimension data, data extraction, and data exhibition. Via multi-dimension data model module, data extraction module, and multi-dimension data visualization based data exhibition module, we provide an effective solution to campus data analysis and visualization. Using the framework, both business analyzers and developers can rapidly develop new data analysis and visualization function.

2. A MDA-based data analysis and visualization framework

In order to solve the problems of existing campus data analysis and visualization, we design and implement a MDA [4] based data acquisition, analysis, and visualization framework. The framework is composed of data visual modeling module, data acquisition module, and visual exhibition module. "Fig. 1" shows the overview design diagram of the proposed framework.

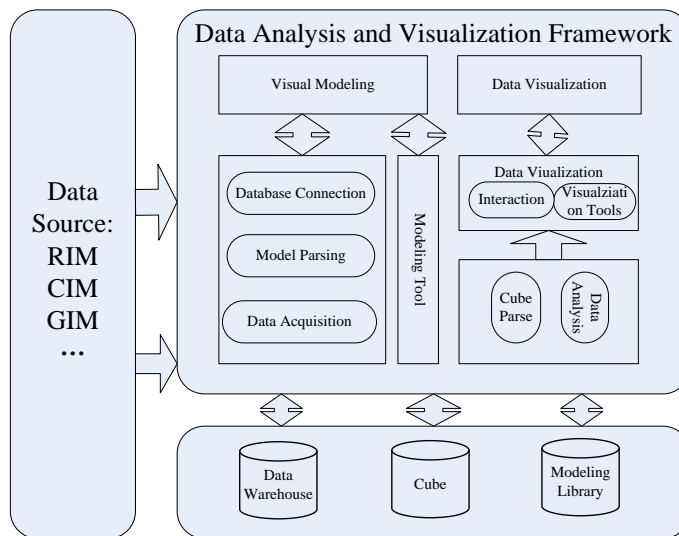


Figure 1. Data Analysis and Visualization Framework

During the process of data analysis and visualization, we first establish the data model by visual modeling related business data, then based on the data model, we conduct data extraction, clean and get all the data sets that we plan to exhibit, finally, we interactively exhibit the model so we can visual exhibit and analysis the data.

A. Data Analysis Modeling

Data model is a typical representation of data structure. According to the requirements of user orientation, we gradually abstract, refine data from the real world, and hierarchically establish data models in different details. These data models are our understanding to the real word in different abstract extents. The data model in traditional business system is relationship data model which cannot effectively express the structure and semantic information between data sets. The primary goal of data analysis system is to analyze a certain topic or some topics. The topic is called fact or measurement, the perspective to do the analysis is called dimension. In this paper, we adopt multi-dimension data model to establish the data model for data analysis and visualization. The system can establish the hierarchical structure of dimensions and measurements, then analyze the tread during a continuous time period, produce data subset slice and rapidly create a new view to show this slice.

The multi-dimension data analysis model integrates both detail information and aggregation information, and satisfies the requirements of decision support system. The model effectively organizes different parts of the system, then produces the aggregation data and summarization data which can provide decision support information to decision maker. The logic structure design of multi-dimension data involves the structure design of fact tables and dimension tables, and the relationships among these tables. According to the specific analysis topic, we first divide the schema into some concrete dimension tables and fact tables, then determine the relationship model for each table, and finally get the logic definition of relation schema.

The multi-dimension data model, organized in the structure form of dimension tables and fact tables, can be represented by star schema, snowflake schema, and fact constellations. The multi-dimension data model usually considers data as a data cube which is composed of dimensions and facts. A dimension is a collection of same type data, each item in the collection is a perspective or an entity that the organization want to record. We adopt the star schema in order to achieve high data query performance and make the model easily understandable. We establish the multi-dimension data model in the following steps.

Step 1: determine the analysis topic. For example, the topic is analyzing every teacher's performance in the past year. This topic is a fact or measurement in the multi-dimension data model.

Step 2: determine the analysis dimensions. In this case, the dimensions include research dimension, teaching dimension, students' supervision dimension, position assessment dimension, and academic reputation dimension. Each dimension can be organized into different levels according to the actual situation, and we will get more small dimensions. It is good for the operations to the data cube, such as slice.

Step 3: based on the analysis in the above steps, we can get a multi-dimension data model which is represented as a star schema.

B. Data Acquisition

For a specific business analysis topic, after completing the multi-dimension data modeling, we extract the data for analysis and visualization from different data sources. In order to provide a clean, complete, accurate and no-repeated data set, improve the analysis performance, and rapidly create the exhibition results, we design an ETL process which extract, transform, and load data from multiple data sources to multi-dimension data set. "Fig.2" shows the data acquisition framework. In the framework, the data from data source is not stored into the multi-dimension data set directly. We firstly pre-process the data. Specifically we transform and clean data according to the corresponding relationship between the source data and the desired data. The transformed data is middle data. Then the middle data undergoes dimension transform and the result data are store into multi-dimension data set.

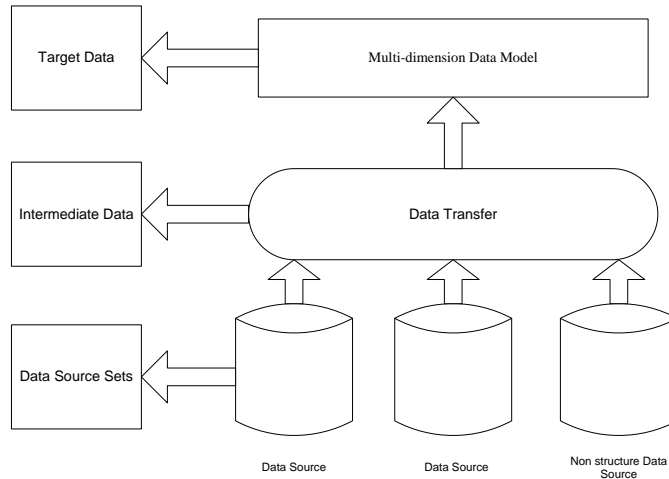


Figure 2. Data Acquisition Framework

As shown in “Fig 2”, the model divides the application system into three layers. They are data source layer, middle data layer, and multi-dimension data layer. According to the features of data, the data in data source layer is divided into structure data and non-structure data. In the paper, we design different data acquisition methods these two data types. For structure data, we adopt traditional data extract, transform, and load process. For non-structure data, we adopt interface-based methods. We use four typical interface types include Web Service, middle database, file, and TCP/UDP packet transfer. The proposed interface-based data acquisition method achieves transparency to heterogeneous data acquisition. Then the heterogeneous data obtained using different interfaces is transformed and loaded into the multi-dimension data set.

C. Data Exhibition

Based on computer graphics and image processing techniques, data visualization converts data to graphics or images, displays graphics or images on digital displays, and provides interaction with the graphics and images. Instead of using relationship tables to exhibit the data, data visualization uses intuitive and effective methods show the relationship among data items [5][6]. After completing the multi-dimension data modeling and data acquisition, we adopt visualization techniques to exhibit data. The common data exhibition methods convert data to a vector art or a bitmap, and display bar chart or/and pie chart using a vector art or a bitmap. For example, some methods visualize data using Scalable Vector Graphics (SVG). SVG is one of the vector graphics standards which uses XML language represents 2D vector graphics, such as rectangle, circle, polygon, and some other geometry graphics. SVG files are of high quality, small size, and rich exhibition effects. However, in order to display SVG, users must install SVG plug-in to their browsers. This may cause inconvenience to users.

In order to make our data exhibition method applicable to all kinds of architecture designs including B/S and C/S, we build our data exhibition model based on Flex techniques. In this way, our data exhibition method can satisfy the display requirements from both B/S and C/S. So users can conveniently use our method. The Flex-based data exhibition process includes data transition, interactive exhibition, and data binding. We obtain the result data from server using the callback mechanism between Flex and Java at the front-end. Flex uses binary “amf” protocol to implement data conversion and binding, so data can be converted between Flex and Java. In the paper, we provide some data exhibition forms for multi-dimension data, including pie chart, bar chart, line chart, parallel coordinate axes [7][8] and scatter chart [9], and so on.

3. Example analysis

Based on the above analysis, we design the data analysis and visualization process as follows. Firstly, we get data model from related business data via visualization modeling. Secondly, according to the data model, we extract and clean all the data sets which are necessary for exhibition. Finally, we interactively determine the data exhibition model and conduct visual exhibition and analysis. We take the example of the number of students to verify the effectiveness of this framework.

In this case, we choose Cube model to represent college-graduation information model. The dimension tables of this data model include college dimension, student type dimension, gender dimension, major dimension, and graduate status dimension. During the implementation of the framework, we use XML technique to model and store the model data. Take the college dimension table for example as shown following XML description, we model and save the following information, attributes, data source table, data choose criteria, group criteria, and etc. After defining the multi-dimension model, we extract and filter data. We extract data using data extraction modeling. Firstly, we use a model to describe data source, data filter criteria, and heterogeneous data acquisition method. Then we analyze the model and extract data. Finally, we store the extracted data as a data table into the multi-dimension model.

```

<dimension name="dimension_college" description=" " >
  <table name="dimension_college" />
  <attributes>
    <attribute name="collegeId" .../>
    .....
  </attributes>
  <sourceTables>
    <sourceTable name="base_college" />
  </sourceTables>
  <wheres>
    <where column="" function="" operator="" />
  </wheres>
</dimension>

```

Figure 3. Multiple Dimension Data XML Description

We provide interactive data exhibition for multi-dimension data based on multi-dimension data visualization components of FLEX techniques. As shown in “Fig4”, we use pie chart to exhibit the numbers of students in different colleges and majors. Firstly, according to business requirements, we choose data dimensions and attributes to exhibit. Then, our system analyzes the data to be exhibited, binds the data with multi-dimension data visualization component, and exhibits the data.

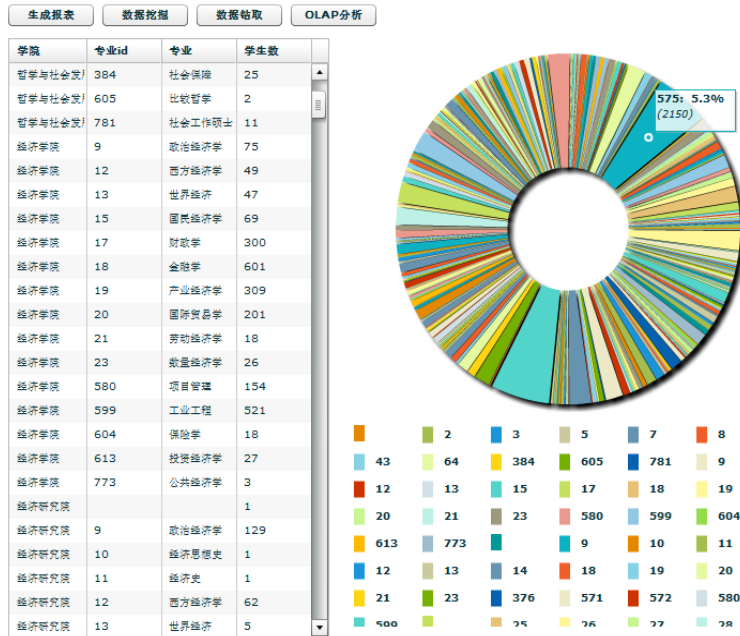


Figure 4. Data Exhibition for Multiple Dimension Data

4. Conclusion

Data analysis and visualization is an emerging and challenging research topic in campus information management. As the rapid development of high education informatization, it will be a new important research topic to conduct the data mining and visual analytics to the existing data in the separated information systems. In order to solve the existing problems in campus data analysis and visualization, we propose a MDA-based complete solution to campus data analysis and visualization. The framework we proposed is composed of multi-dimension data modeling module, data extraction and clean module, visualization-based data exhibition module. Using the MDA-based framework, both business analyzers and developers can rapidly implement the modeling and analysis process for business data visualization. The framework also provides interface with data mining. In the future, we will explore data mining algorithms [10, 11, 12] and multi-dimension data visualization techniques, and integrate these algorithms and techniques into our data analysis and visualization framework.

References

[1] G. Bastien, Sen Wu, Xuedong Gao. "Data warehousing and data mining", China Machine Press, 2001, pp. 55-100.

[2] C. Y. Ren, W. Q. Luo,. "The Combination and Application of CRM and DM in Campus Information System"[J]. Computer Engineering and Applications, 2003. (In Chinese)

[3] J. Li, "Data Mining and Its Application in Campus Teaching Informatization Management" [J]. Colleges and Universities Science Research, 2007. (In Chinese)

[4] Model Driven Architecture , <http://www.omg.org/mda/index.html>.

[5] Heer, J., S. K. Card, J. A. Landay. prefuse: A Toolkit for Interactive Information Visualization. ACM Human Factors in Computing Systems (CHI), 2005.

[6] SUN Yang, FENG Xiaosheng. "Survey on the Research of Multidimensional and Multivariate Data Visualization", Computer science, 2008, pp.1-2.

- [7] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In Proc. of IEEE Visualization, pages 43–50, 1999.
- [8] K. T. McDonnell and K. Mueller. Illustrative parallel coordinates. Eurographics/IEEE-VGTC Symposium on Visualization. Vol 27(3), 2008.
- [9] Sven Bachthaler and Daniel Weiskopf. Continuous scatter plots. IEEE Trans. on Vis. and Comp. Graph., Vol.14(6), pages 1428-1435,2008.
- [10]J. LeBlanc, M.Ward, and N.Wittels. Exploring n-dimensional databases. Proc. of Visualization '90, p. 230-7, 1990.
- [11]T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. SIGMOD Record, vol.25(2), p. 103-14, June 1996.
- [12]YANG Dongqing, MA Xiuli and TANG Shiwei. "Business Modeling and Data Mining", China Machine Press, 2005, pp.56-73.