Application Research of Data Warehouse Technology in College Student Enrollment and Employment Decision

Shen Zhang [a], Yuliang Shi [b]

Faculty of information technology, Beijing University of Technology, Beijing 100124, China.

[a] maggiepipixia@163.com, [b] shiyl@bjut.edu.cn

Abstract. Data warehousing is an emerging technology that has been developed on the basis of database technology. A data warehouse is a topic-oriented, integrated, non-volatile, time-varying data set used to support executive decision making. To some extent, students are products produced by schools, and characteristics of students are at school. The high registration rate and the characteristics of the students in which departments have high employment rate are all very concerned about the school. Which schools and regions are enrolled in employment or employment situation, which is the higher priority of the higher authorities and school leaders. The problem. This topic is to establish a data warehouse for enrollment and employment in colleges and universities. Through the multi-dimensional analysis of data warehouses, we can discover the hidden rules behind the data and discover the inherently meaningful knowledge, so as to better guide the management and decision-making of enrollment and employment. Enrollment and employment decision-making require a large number of multi-sources, heterogeneous, and dynamic data. Data warehouse technology is applied in the field of enrollment and employment. Various potential and valuable rules are found from many historical data to scientifically guide employment. Efficiently carrying out publicity will help to improve the quality of college enrollment and employment, as well as the quality of graduate employment. It has important practical significance for the development of the entire university and the improvement of the quality of students.

Keywords: Data warehouse; decision support system; college admissions employment.

1. Introduction

As the number of college entrance examinations declines year by year, colleges and universities face severe enrollment problems every year. Especially how vocational colleges survive in fierce competition, employment rate and employment direction, employment salary is a very important indicator. Now every college has made detailed statistics on its annual total employment rate, employment rate and employment direction of different professions. Therefore, a large number of graduate employment information has been accumulated in the database. These data are valuable resources for colleges and universities. How to use these data to make decision analysis on a macro level, reflecting its true value. Through the analysis of existing data, knowledge extraction finds the degree of relevance between profession and employment, employment area and unit nature. Through the extraction of knowledge, it can be used for professional construction, career goals, career direction, course orientation, etc. There is considerable guidance on how to make a technical guidance or how to set up a professional, reasonable and efficient curriculum for the market.

As a new and multidisciplinary application field, data mining is playing an increasingly important role in the information analysis-based decision support system activities of all walks of life. As an important research branch of data mining, data warehouse mining is mainly used to discover related relationships between items in data sets, that is, association rules. Data mining is the main means of extracting knowledge information from large databases. Because of its simple form and easy understanding, data warehouse mining has been widely used in various fields to test long-term knowledge patterns in the industry or to discover hidden new laws.

Existing decision tree algorithms also have many shortcomings, such as multi-value bias of attribute selection and attribute vacancies.

Processing of values, processing of continuous values of attributes, etc. So how to further improve the performance of the decision tree and improve its

Classification accuracy, making it more suitable for data mining applications requires important theoretical research significance and reality significance. This paper conducts in-depth research on the inadequacies of the above decision trees and explores the decision tree classification algorithm.

Optimization algorithm and how to use the decision tree method to classify the graduate data warehouse.

$$\mathrm{B}(A)=\sum_{j=1}^{v}\sum_{i=1}^{k}\frac{|s_{ij}|(|s_j|-|si_j|)}{|s_j|}$$

2. Implementation of Data Warehouse for Enrollment and Employment Decision

2.1 Source Data Preparation

The creation of the data warehouse mainly includes four core steps of data extraction (Txtraction), data transformation (Transformation), data cleaning (Cleansing) and data loading (Loading), which is the most difficult and most difficult part of the data warehouse construction process. The ability to integrate and improve the value of data according to uniform rules, responsible for transferring relevant data to the fact table and dimension table of the data warehouse, and completing the process of converting data from source data to data warehouse is an important step in implementing data warehouse. The main tasks of ETL include:

(1) Identification and extraction of relevant information on the source data side;

(2) Develop and integrate information from multiple data sources into a unified format;

(3) on the basis of the database and business rules, clean up the extracted result data set;

(4) Data is transmitted to the data mart or data warehouse

2.2 Data Warehouse Creation and ETL

Since the current mainstream data warehouse solutions provide ETL tools, the ETL tools that come with the data warehouse platform can solve the problem of different data source integration and integration at a relatively low cost. For example, DTS in SQL Server is a powerful set. Data conversion tool for data import, export, and conversion services between SQL Server and other data sources (OLEDB, ODBC) and text files, spreadsheets. With DTS, data warehouses and data marts are easily built on SQL Server by interactively entering data from multiple heterogeneous data sources on an interactive basis.

2.2.1 Ways to Achieve

Since the current mainstream data warehouse solutions provide ETL tools, the ETL tools that come with the data warehouse platform can solve the problem of different data source integration and integration at a relatively low cost. For example, DTS in SQL Server is a powerful set. Data conversion tool for data import, export, and conversion services between SQL Server and other data sources (OLEDB, ODBC) and text files, spreadsheets. With DTS, data warehouses and data marts are easily built on SQL Server by interactively entering data from multiple heterogeneous data sources on an interactive basis.

2.2.2 Concrete Implementation

create a database, use DTS for data conversion and loading

The fact tables and dimension tables in the data warehouse are organized according to the employment information of students, but they are stored in a traditional database or spreadsheet in the form of two-dimensional tables. Therefore, these data should be used by OLAP tools. Multidimensional processing to form a multidimensional data model (data cube).

After the data is pre-processed, the employment data from 2007 to 2011 (provided by EXCEL) is aggregated to form a summary of employment information. Create a new database, use DTS to load data into the database.

2.3 Implementation of OLAP

Once the data warehouse is created, you can start Analysis Services, establish a connection to the data warehouse created above, and create a cube.

The Analysis Services integrated environment provides valuable insights into multidimensional analysis of cubes, such as data browsing, drilling, scrolling, slicing, and dicing. The data is analyzed in terms of “professional” and “city”. The professional can subdivide different cities below, and the fields with ‘+’ can continue to be subdivided, that is, “drill down”. The analysis can show the different distribution of different professions in the employment city, can draw the connection between professional and employment cities, get the employment rate of some professions in which cities and urban areas, and which cities are the type of professional employment rate is relatively low, and it is very intuitive. For example, the software technology major in Hainan followed Shenzhen's employment rate is relatively high, while in the Northeast and Hong Kong, Macao and Taiwan regions are almost empty; education majors, such as primary education, music education, modern education technology is almost employed in Hainan, other The areas are relatively small; in general, computer-related majors are concentrated in the Central and South regions, followed by the East China region.

3. Algorithm Research and Algorithm Improvement

It can be seen from the basic formula of information gain that the size of the information gain is determined by the information entropy of the attribute. Information entropy is used to reflect the degree of uncertainty of each attribute to the entire data sample set [31]. This paper balances the uncertainty of each attribute on the data sample set by introducing a weight n to the information entropy of each attribute, making it more in line with the actual data distribution. For the introduction of weights, this paper uses the number of values of each attribute in the data sample set, and multiplies the value of the information by the weight n, so that the information entropy result depends on the value of the attribute. It is to take into account the value of each attribute, thus overcoming the multi-value bias problem when the ID3 and C4.5 algorithms select test attributes.

References are cited in the text just by square brackets [1]. (If square brackets are not available, slashes may be used instead, e.g. /2/.) Two or more references at a time may be put in one set of brackets [3, 4]. The references are to be numbered in the order in which they are cited in the text and are to be listed at the end of the contribution under a heading References, see our example below.

According to the optimization of the attribute rule, E(A) can be simplified to

$$\mathrm{B}(A)=\sum_{j=1}^{v}\sum_{i=1}^{k}\frac{|s_{ij}|(|s_j|-|si_j|)}{|s_j|}$$

In order to avoid the multi-value bias problem of attribute selection, we can substitute the weight n, then the new simplified entropy B'(A) can be simplified to

$$\mathrm{B}'(A)=n\sum_{j=1}^{v}\sum_{i=1}^{k}\frac{|s_{ij}|(|s_j|-|s_{ij}|)}{|s_j|}$$

4. Summary

(1) There are great differences in the employment rate of different majors in colleges and universities. Some professional employment rates are as high as 98%, and some professional employment rates are only 20%. There is a huge correlation between the employment rate of different majors and the demand for social talents. In the process of reviewing the school enrollment plan in the follow-up schools, the number of plans is appropriately changed in accordance with the demand for social talents to meet the needs of the society.

(2) Through the analysis of employment rates at different levels of education, it is found that the employment rate of undergraduates is 94%, the employment rate of graduate students is 61%, the employment rate of doctoral students is 68%, and the employment rate of junior college students is 84%. According to the above data, it has certain reference significance for graduates who are interested in pursuing high education.

(3) Through the analysis of graduate employment units, we can find that there is a huge correlation between the graduate level of the graduates and the nature of the employment units. In recent years, a series of national implementation plans, such as three support, one village student, and the western plan, are the primary choices for graduates.

(4) Graduates of undergraduate degree are more favored in developed provinces and cities, such as Beijing, Shanghai, Shenzhen, etc. in the process of employment enterprises and employment selection. Unless there are special circumstances, graduates will choose some third- and fourth-tier cities to choose employment. Although this employment situation is in line with the development trend of the social economy, it is a fatal blow to the overall balanced development of the country. Taking effective measures to attract graduates to come to support is a problem that governments in various regions must solve.

References

[1]. Han J, Kamber M. Data Ming: Concepts and Techniques. [M]. San Francisco: Morgan Kaufmann Publishers.2001.

[2]. Clemen R T. Making Hard Decisions with Decision Tools. [M]. Belmont: Duxbury, 2000.

[3]. X. Chen, J. Li, W. Zhang, Z. Zou, F. Ding, Y. Liu and Q. Li. The Development and Application of Data Warehouse and Data Mining in Aluminum Electrolysis Control Systems. Light Metals 2006.Proceedings of the Technical Sessions Presented by the TMS Aluminum Committee at the 135th TMS Annual Meeting,12-16 March 2006, Warrendale, PA, USA ,2006. TMS (The Minerals, Metals & Materials Society).2006.

[4]. Xuejian Yan, Xueqing Li. A Multidimensional Data Analysis System Based on MDA for Educational Data Warehousing. The 6th International Conference on Computer Science & Education. Singapore, August 3,2011.

[5]. K. W. Chau, C. Ying, M. Anson and Z. Jianping. Application of Data Warehouse and Decision Support System in Construction Management. Automation in Construction.2003,12(3):213-224.

[6]. Kamran Parsaye. Surveying decision support[J]. Data programming and design,1996,9(4):27-33.

[7]. TungX, Bui. decision support in the future tense[J]. Decision Support System,1997, (19):85-87.

[8]. Q. Hanand Z. He. Research on Cost Control Dss Based on Knowledge Warehouse.6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009, August 14,2009-August 16,2009, Tianjin, China, 2009.IEEE Computer Society,2009:357-361.