A DATA MINING ANALYSIS OF COLLEGE ENGLISH TEST (CET) RESULTS OF HAOJING COLLEGE, CHINA

CHEN DONG

A dissertation submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Information Technology - Management)

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2013

To my beloved mother and father

ACKNOWLEDGEMENT

First of all, I want to thank my supervisor Dr. Alex Sim for valuable advices, guidance, and encouragement. I received thoughtful suggestion and warm spiritual support from all of postgraduate students in the field of IT management. I would like to thank all of my friends from especially Wang Xinxin, Jason, George, Coco and also Vivian for their support. I also need to thank my dear parents who give more love and encouragement.

ABSTRACT

The data mining technology is one of the hot issues of information technology research. Data mining has already used in many fields, such as bank, finance, insurance, and retail. But unfortunately, data mining technology is seldom used in the field of education. With the continuous enrollment of Chinese universities, more and more students go into university to study. Meanwhile, a mass of data were produced which about the students' basic information and their subject mark. There are still some deeply relationship among subject mark under cover. Using data mining technology into analysis of students' mark can find the real factors which affect their subject and then improve the teaching quality. In this paper, the theoretical knowledge of data mining was studied. Research and Application of data mining technology in student CET mark analysis is based on data mining technology research. Before the excavation, the paper established students' achievement analysis of data, and data cleaning, data conversion, data reduction, data preprocessing, processing vacancy data, the continuous-valued attribute discretization, to lay the foundation for further excavation. After comparison of various algorithms in data mining, association rules algorithm was chosen, which is suitable for student achievement analysis model Apriori algorithm to conduct students' CET mark analysis. In the final realization of the process, the relationship between CET mark and employment salary were found and also found the main factors which affect the CET mark. These finding are very useful which can improve the teaching quality.

ABSTRAK

Teknologi perlombongan data adalah salah satu daripada isu-isu panas penyelidikan teknologi maklumat. Perlombongan data telah digunakan dalam banyak bidang, seperti bank, kewangan, insurans, dan runcit. Tetapi malangnya, teknologi perlombongan data jarang digunakan dalam bidang pendidikan. Dengan pendaftaran berterusan universiti China, semakin ramai pelajar pergi ke universiti untuk belajar. Masih terdapat beberapa mendalam hubungan antara tanda tertakluk di bawah perlindungan. Menggunakan teknologi perlombongan data dalam analisis cap pelajar dapat mencari faktor-faktor sebenar yang memberi kesan kepada subjek mereka dan kemudian meningkatkan kualiti pengajaran. Dalam kertas ini, pengetahuan teori perlombongan data telah dikaji. Penyelidikan dan Penggunaan teknologi perlombongan data dalam pelajar CET analisis markah adalah berdasarkan perlombongan data penyelidikan teknologi. pemprosesan data kekosongan, yang pendiskretan sifat berterusan-bernilai, untuk meletakkan asas bagi penggalian selanjutnya. Selepas perbandingan pelbagai algoritma dalam perlombongan data, persatuan peraturan algoritma telah dipilih, yang sesuai untuk pelajar pencapaian analisis model Apriori algoritma untuk menjalankan CET analisis markah pelajar. Dalam merealisasikan akhir proses itu, hubungan antara CET tanda dan gaji pekerjaan telah ditemui dan juga mendapati faktor-faktor utama yang memberi kesan kepada tanda CET. Penemuan ini adalah sangat berguna yang boleh meningkatkan kualiti pengajaran.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

In this chapter, the study background information, statement of problems, objectives, research questions, significance of study, and scope of study will be mentioned.

1.1 Background Information

With the continuous enrollment of Chinese universities, more and more students go into the University to study; the numbers of university students have reached several thousand or even tens of thousands. However, with the growth in the number, it also brought large number of the data, which are about students' basic information and their performance.

In Chinese college, there are some ways to measure the students' capability. For example, if any student wants to get the Bachelor Degree Certificate, they must pass the CET-4. CET-4 means College English Test level 4, this certificate is the normal way to measure students' English level. When the students pass the exam, that is means the student has normal English skill. And China also has the CET-6 test,

which is means College English Test level 6. CET-6 is higher level English certificate, when the students pass the test, it means the student know the English very well.

Last but not the least is that China is very large, all the students in college come from different district, different district have the different admission scores to go to the college.

In every Chinese college, they have all the data about the students' academic marks, CET-4 and CET-6 marks and the students' native place information. The college also will record which job the students acquired after their graduation. So face the so large students' data, what should the college do to analysis the data and whether there have any relationship between the students' information that is a problem need to solve.

1.2 Statement of problems

Data mining has already used in many fields, such as bank, finance, insurance, and retail. Used data mining, these fields can based on the result of data mining, discover the potential customer and change their strategy. But unfortunately, data mining technology is seldom used in the field of education. Every Chinese institution has large data about students, but the college never found the relationship between different data.

Nowadays, in nearly every Chinese college, the data about the students will be recorded in the Excel, and different department kept the different data. For

example, all the marks of students' academic subjects will be preserved by the department of each faculty. However, the department of English will preserve the marks about the CET-4 and CET-6, the information of obtain employment of students will be seen in the admission and employment office. Therefore, it is difficult for college staff to compare the data and find the relationship between the different data.

On the other hand, in Chinese colleges, each semester, the students will be given more than eight subjects, different field students have the different subjects, but every subject needs a middle test and a final exam, so after four years undergraduate study, they have lots of marks about their study. Besides this, after each semester, every faculty will calculate the GPA (Grade Point Average) for students.

Besides, the use of data still remains at level of transaction management and information retrieval (increase, delete, modify, inquire, statistics), But cannot find the relationship that exists in the data and rules, cannot predict the future development trend based on available data. Actually, according to the students' data, it includes large useful information beside these students' data, and this information can provide the basis for decision making to college and education.

1.3 Objectives

(1) To study the influences of College English Test on students' salary after graduation.

(2) To identify the main factors which affect the result of College English Test.

(3) To recommend strategies for improving English teaching quality in a China college.

1.4 Research Questions

(1) What are influences of College English Test on students' salary after graduation?

(2) What are the main factors which affect the result of College English Test?

(3) What recommendations can be given to strategies for improving English teaching quality in a China college?

1.5 Significance of the Study

As the author mention before, with the rapid development of Chinese economy, and after China joined WTO (World Trade Organization) in 2006, more and more famous foreign companies go to China to establish their Branch Office. Therefore, English became more and more important for students if they want to acquire a good job after their graduation.

CET is a national English test which is means College English Test. It is provided by Chinese Education Department, the students have two chances to take the exam in one year, which is on June and December. CET 4 is a lower level English test compare with CET6. Most of the colleges provide that if the students who want to get the bachelor degree, they must pass the CET 4 test. CET 6 means College English Test 6, which is a higher level English test, if the students pass the exam, it can be said that the students have good English skill.

Nowadays, when the students to employ the companies, nearly all the companies ask students to show the CET certificate, only the students have the CET certificate, the companies receive the students' resume and most of the famous

companies requires that the students must pass CET 6, they can accept the resume and estimate whether the students fit for this company.

Thus, using data mining technique, the college can find the relationship and rules between students' CET mark and situation of employment, and then help them improve their teaching quality and give the students some suggestion when they find job. Besides, the college can use the result to guide their strategy.

Nowadays, China has lots of colleges, but seldom used data mining to improve their teaching quality and guide their strategy. So this thesis has large value of research, if use the data mining technique, the school gets a good result, this will be a good news for Chinese college students and educators, meanwhile the college can solve the problem of wait for employment.

1.6 Scope of the study

In this thesis, data will be collected from Haojing College of Shaanxi University of Science and Technology in China.

Haojing College of Shaanxi University of Science and Technology is a new college in China. It is founded on 2004. The college always adheres to the correct guiding ideology and educational philosophy, the school motto is "Simple and efficient, the pursuit of excellence". From 2004 to 2012, the students number in the school from 400 to 8,000.

The author will use the data from 2004 to 2011 about students' basic information, and their marks on their academic subjects, their CET-4 mark, CET-6 mark, and the data about their employment information and also their job situation.

The entire student's data will be collected from different faculty in Haojing College of Shaanxi University of Science and Technology and the majors include International Economy and Trade, Information Management and Information Systems, Human Resource Management, Administration, Marketing, Accounting, Computer Science and Technology, Logistics Engineering, Fashion Design and Engineering, Electronic Science and Technology, Electrical Engineering and Automation, Electronic and Information Engineering, and Network Engineering.

REFERENCE

Bayram. (2010). Data mining application on students′ data. . *Procedia Soc Behav Sci*, 5251－5259.

Caifang, Y. (2011). Improvement of English reading in University. *Foreign language and study*.

Chang-xin, S., & Ke, M. (2008, 20-22 Dec. 2008). *Applications of Data Mining in the Education Resource Based on XML*. Paper presented at the Advanced Computer Theory and Engineering, 2008. ICACTE '08. International Conference on.

Feng, Z. (2010). The Study of Spoken English Teaching in Chinese College. *Read and Write Periodical, Vol. 6*.

Garc á. (2010). collaborative educational association rule mining tool，Internet and Higher Education.

Hamalainen. (2008). Data mining personalizing distance education courses. *World conference on open learning and distance education，Hong Kong*.

Hanjun, J., & Tianzhen, W. (2009, 7-8 March 2009). *Application of Visual Data Mining in Higher-Education Evaluation System.* Paper presented at the Education Technology and Computer Science, 2009. ETCS 09. First International Workshop on.

Huo, S. (2009). Association rule mining theory in teaching evaluation. *Software GUIDE*, 11-12.

Jia-Lang, S. (2010). An analytic approach to select data mining for business decision. [doi: 10.1016/j.eswa.2010.05.083]. *Expert Systems with Applications, 37*(12), 8042-8057.

Ling, J. (2008). Summarization on the Data Mining Application Research in Chinese Education Advances in Blended Learning. In E. Leung, F. Wang, L. Miao, J. Zhao & J. He (Eds.), (Vol. 5328, pp. 110-120): Springer Berlin / Heidelberg.

Ngai, & Hu, Y. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. [doi: 10.1016/j.dss.2010.08.006]. *Decision Support Systems, 50*(3), 559-569.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. [doi: 10.1016/j.dss.2010.08.006]. *Decision Support Systems, 50*(3), 559-569.

Peng, J. (2012). DIsscuse the reformation of English education in Chinese college. [Education]. *China science and technology information*, 10.

Romero, & Crist óbal. (2008). Data mining in course management systems: Moodle case study and tutorial. [doi: 10.1016/j.compedu.2007.05.016]. *Computers & Education, 51*(1), 368-384.

Shahriar, M. S., & Anam, S. (2008, 13-15 Dec. 2008). *Quality Data for Data Mining and Data Mining for Quality Data: A Constraint Based Approach in XML*. Paper presented at the Future Generation Communication and Networking Symposia, 2008. FGCNS 08. Second International Conference on.

Sharma, S. (2012). Evaluation of an integrated Knowledge Discovery and Data Mining process model. [doi: 10.1016/j.eswa.2012.02.044]. *Expert Systems with Applications, 39*(13), 11335-11348.

Taichang, H. (2008). Course planning of extension education to meet market demand by using data mining techniques — an example of Chinkuo technology university in Taiwan. *Expert Systems with Applications*, 596—602.

Talavera. (2009). Mining student data to chaacterize similar behavior groups in unstructured collaboration spaces. *European conference on artificial intelligence*, 17-23.

Xiao, C. (2010). Data Mining Research and Application of K-MEANS algorithm in the students CET4 performance analysis. *Journal of shanghai institute of tec*

hnology.

Yan, Z. (2010). Improvement of English teaching in college. Education. *Jiangxi exam weekly magazine*, 8.

Yaqin, F. (2010, 14-15 Aug. 2010). *XML in Web Data Mining Application.* Paper presented at the Information Engineering (ICIE), 2010 WASE International Conference on.

Zhengdong, Z. (2012). English dictation study in Chinese college. *Education and study*.

Zhihua, Z. (2011). Disscuss the reformation of Chinese English education by comparing the distinguish of Amrican education and Chinese. *China science and technology information*.

Zhimin, W. (2009). Application ofOLAP and Data—mining inThe VisualPrediction System of Passing Rate in CET‘4 andCET‘6. *Journal of shanghai institute of tec hnology*.