MINING SEBHA UNIVERSITY STUDENT ENROLMENT DATA USING DESCRIPTIVE AND PREDICTIVE APPROACH

A thesis submitted to the Faculty of Information Technology in partial Fulfillment of the requirement for the degree Master of Science (Intelligent System) Universiti Utara Malaysia

By
MANSOUR ALI ABDOULHA

KOLEJ SASTERA DAN SAINS
(College of Arts and Sciences)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

MANSOUR ALI ABDOULHA
(800301)

calon untuk Ijazah
(candidate for the degree of) **MSc. (Intelligent System)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/her project paper of the following title)

MINING SEBHA UNIVERSITY STUDENT ENROLMENT DATA USING DISCRIPTIVE AND PREDICTIVE APPROACH

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory knowledge of the field is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **ASSOC. PROF. FADZILAH SIRAJ**

Tandatangan
(Signature) : Fadzilah

Tarikh
(Date) : 18/11/2008

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor, in her absence, by the Dean of the Faculty of Information Technology. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain should not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Request for permission to copy or to make use of material in this thesis, in whole or in part should be addressed to:

Dean of the Faculty of Information Technology

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Darul Aman

ABSTRAK
(BAHASA MELAYU)

Salah satu perkara yang dititikberatkan oleh sistem pengajian tinggi ialah dalam menilai dan menambahbaik organisasi pendidikan. Untuk mencapai objektif, sesebuah organisasi itu perlu mempunyai pengetahuan mendalam untuk mencapai, menilai dan membuat perancangan ke arah proses pembuatan pemutusan yang lebih baik. Teknik Perlombongan Data merupakan alat bantu analisis yang boleh digunakan untuk menperoleh pengetahuan dari pangkalan data yang besar. Kajian ini membentangkan dapatan kajian penggunaan Perlombongan Data dalam bidang pengajian tinggi, terutamanya bagi Universiti Sebha, Libya. Sumbangan utama kajian ini ialah satu model analisis yang boleh digunakan sebagai alat bantu pembuatan pemutusan. Ianya berfungsi, sebagai garis panduan untuk mengenalpasti bahagian yang boleh dipertingkatkan melalui teknologi Perlombongan Data serta bagaimana teknologi tersebut menambahbaik proses konvensional. Dua pendekatan yang digunakan, pertama, kaedah ***descriptive statistic*** *terutamanya* ***cross tabulation analysis*** *digunakan bagi menerangkan maklumat yang terdapat dalam data. Analisis pengelompokan telah digunakan untuk mengkelaskan data ke dalam beberapa kelompok berdasarkan kesamaan. Kluster tersebut digunakan sebagai target bagi ujikaji peramalan. Bagi Analisis Peramalan, tiga teknik telah digunakan iaitu Rangkaian Neural, Regresi Logistik dan Pohon Pemutusan. Dapatan kajian menunjukkan Rangkaian Neural memperolehi ketepatan tertinggi berbanding dua teknik Regresi Logistik dan Pohon Pemutusan.*

ABSTRACT
(ENGLISH)

One of the main concerns of higher educational system is evaluating and enhancing the educational organization. For achieving this quality objective the organizations need deep knowledge assess, evaluate and plan towards better decision making process. Data mining techniques are analysis tools that can be used to extract meaningful knowledge from large databases. This study presents applying data mining in the field of higher educational especially for Sebha University in Libya. The main contribution of the study is an analysis model that can be used as a decision support tool. It acts as a guideline or roadmap to identify which part of the processes can be enhanced through data mining technology and how the technology could improve the conventional processes by getting advantages of it. Two main approaches were used in this study. Firstly the descriptive statistics, particularly cross tabulation analysis was carried out and presents a lot of useful information within data. Cluster analysis was performed to group the data into clusters based on its similarities. The clusters were also used as targets for prediction experiment. For predictive analysis, three techniques have been used Neural Network, Logistic regression and the Decision Tree. The study shows that Neural Network obtains the highest results accuracy among the three techniques.

ACKNOWLEDGEMENT

In the name of Allah, Most Gracious, Most Merciful. Peace upon the prophet, Muhammad S.A.W, a foremost praise and thankful to Allah for His blessing, giving me the strength in completing this research.

I would like to extend my thanks and gratitude to my supervisor Associate Professor Fadzilah Siraj for the guidance, patience, encouragements, advice and flourish of knowledge during completing this research.

Last but not least, a lasting heartfelt to my wife for her patient during MSc period, and to my new porn son Ali, also great thankful to my mother and father and family in Libya for all their love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

Data collection was a very expensive activity in time and resources. Due to the advance in computing and the availability of Internet, this activity had become much cheaper and easier to undertake. Data has become so plentiful that corporation has created data warehouses to store them and hired statisticians to analyze their information. Administrators of student registrations office would make important use of data to inform their registrar's office practice if such information can be easily available to them in a way that fit their needs and answer their questions.

Nowadays, higher educational organizations are placed in a very high competitive environment and are aiming to get more competitive advantages over the other business competitors. These organizations should improve the quality of their services and satisfy their customers (industry, government). They believe that their students and professors are the main assets and they want to improve their key process indicators by effective and efficient use of their assets. To remain competitiveness among educational filed, these

The contents of the thesis is for internal user only

REFERENCES

Anders, K. (2001). *Data mining for automated GIS data collection.* Whchmann veriag, 263-270.

Armstrong, J., & Anthes, K. (2001). *How data can help.* American School Board Journal 188(11), 38-41.

Beikzadeh. M. R., & Delavari. N. (2004). *A New Analysis Model for Data Mining Processes in Higher Educational Systems.* In MMU International Symposium on Information and Communications Technologies 2004 in Conjunction with the 5th National Conference on Telecommunication Technology 2004. Putrajaya,Malaysia..

Cahlink, G. (2004). *Data Mining Taps the Trends.* Government Executive Magazine, Retrived on 2008-07-20, from http://www.govexec.com/tech/articles/1000managetech.htm.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Thomas, R., Shearer, C., & Wirth, R., (2000). *CRISP-DM 1.0 Step-by-step data mining guide.* SPSS White paper-technical report CRISPWP-0800, SPSS Inc.

Chamillard, A.T. (2006) Using Student Performance Predictions in a Computer Science Curriculum. ITiCSE'06, June 26–28, 2006, Bologna, Italy.

Chang, H.C. and Hsu, C.C. (2005) Using Topic Keyword Clusters for automatic Document Clustering. Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05).

Chrispeels, J. H., Brown, J. H. & Castillo, S. (2000). *School Leadership Teams: Factors that influence their development and effectiveness.* Understanding Schools as Intelligent Systems, Vol. 4, 39-73, JAI Press.

Dass, R. (2006). *Data mining in banking and finance: A note for bankers*. Indian Institute of Management Ahmedabad.

Dunham, M.H. (2003). *Data mining introductory and advanced topics*. Upper Saddle River, NJ: Pearson Education, Inc.

Edelstein, H. (1997). Data mining: Exploring the hidden trends in your data.

Efraim, T., Jay, E.A.,Tin-Peng, L. & Ramesh, S. (2007) *Decision Support and Business Intelligent Systems (Eight Edition),* Pearson Education, Inc.

Everitt, B. S. (1993). Cluster analysis (2nd. Ed.). London : Edward Arnold.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery in databases*. American Association for Artificial Intelligence.

Feldman, J., & Tung, R. (2001). *Using data-based inquiry and decision making to improve instruction.* ERS Spectrum 19(3), 10-19.

Goebel. M., & Gruenwald. L. (1999). *A survey of data mining and knowledge discovery software tools*. SIGKDD Explorations ACM SIGKDD 1(1), page 20.

Han, J., & Kamber, M. (2001). *Data mining: concepts and techniques (Morgan-Kaufman Series of Data Management Systems).* San Diego: Academic Press.

Hanna, M. (2004). Data mining in the e-learning domain. on Journal Campus-Wide Information Systems.

Hofmann. M. (2003). *The development of a generic data mining life cycle.* International Conference on Software Engineering Theory and Practice. Proceedings, Orlando, USA, July 9-11.

Kennedy, E. (2003). *Raising test scores for all students: An administrator's guide to improving standardized test performance*. Thousand Oaks, CA: Corwin Press. Retrieved on 2008-07-24 from http://findarticles.com/p/articles/mi_m0JSD/is_8_61/ai_n6191437.

Kusiak, A., Kernstine, K.H., Kern, J.A., McLaughlin, K.A., & Tseng, T.L. (2000). Data Mining: Medical and Engineering Case Studies. *Proceedings of the Industrial Engineering Research 2000 Conference,* Cleveland, Ohio, May21-23, 2000, pp. 1-7.

Laxman, S., & Sastry, P. S. (2006). *A survey of temporal data mining.* Proceedings in Engineering Sciences, Bangalore, India. *31*(2) pp. 173–198.

Luan. J. (2001). *Data Mining as Driven by Knowledge Management in Higher Education-Persistence Clustering and Prediction.* Keynote for SPSS Public Conference, UCSF.

Luan, J. (2002). Data Mining Application in Higher Education. *SPSS Executive Report.* Retrieve from http://www.crisp-dm.org/CRISPWP-0800.pdf.

Luan, J. (2004). Data Mining and Knowledge Management in higher Education Potential Application. *Proceedings of Air Forum, Toronto, Canada.*

Lovis, C. Colaert D, & Stroetmann V. N. (2008). *DebugIT for Patient Safety Improving the Treatment with Antibiotics through Multimedia Data Mining of Heterogeneous Clinical Data.* eHealth Beyond the Horizon – Get IT There.

MacQueen, (1967). *A Tutorial on Clustering Algorithm*. Retrieved from http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html, on 2007.

Marquez, L., Hill, T., Worthley, R., & Remus, W. (1991). *Neural network models as an alternative to regression. System Sciences*, Proceedings of the Twenty-Fourth Annual Hawaii International Conference on Volume iv, 8-11 Jan. 1991 Page(s):129 - 135 vol.4

Netz, A., Chaudhuri, S., Bernhardt, J., & Fayyad, U. (2000). *Integration of Data Mining and Relational Databases*. Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt.

Quinlan. J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publisher: NY.

Romeu, J. L. (n.d.). *Operation research/statistics techniques: A key to quantitative data mining*. IIT Research Institute, Rome, NY.

Rubenking, N. (2001) Hidden Messages. PC Magazine. May 22, 2001. Retrieved on 11 September 2008, from www.pcmag.com/article2/0,2817,8637,00.asp.

Seifert, J. W. (2004). *Data mining: An overview*. Congressional Research Service, the Library of Congress.

Shi. H. (2006). *Best-first Decision Tree Learning*. Hamilton, NZ. M.Sc. Thesis: University of Waikato.

Sinivirta. J. (2006). *Implementing analytics for a rating engine*. M.Sc. Thesis: Helsinki University of Technology, Finland.

Sirikulvadhana. S. (2002). *Data mining as a financial auditing tool*. M.Sc. (Thesis in Accounting). The Swedish School of Economics and Business

Administration. retrived on 12/9/2008, from www.pafis.shh.fi/graduates/supsir01.pdf.

StatSoft White Paper (2007). *What is Data Mining, and How is it Useful for Power Plant Optimization?* Website www.statsoft.com/support/whitepapers/pdf/WhatIsDataMining.pdf.

Stephens, S., & Pablo, T. (2003). *Supervised and unsupervised data mining techniques for the life sciences*. Oracle and Whitehead Institute/MIT , USA

Therling. K (1995) *An Overview of Data Mining at Dun and Bradstreet*. DIG White Paper.

Thorn, C. A. (2001). *Knowledge Management for Educational Information Systems: What is the State in the Field?*. Education Policy Analysis Archives, *9*(47), retrieved July 22, 2008, from http://epaa.asu.edu/epaa/v9n47/.

Toprak. S. (2004). *Data mining for rule discovery in relational databases*. M.Sc. Thesis: Middle East Technical University.

Two Crows Corporation (1999). *Introduction to Data Mining and Knowledge Discovery*, 3ed ed. Potomac, MD: Two Crows Corporation.

Tsantis. L., & Castellani. J. (2001). *Enhancing Learning Environments through Solution-based Knowledge Discovery Tools: Forecasting for Self-perpetuating Systemic Reform.* Journal of Special Education Technology, Volume 16(4).

U.S. General Accounting Office (2004). *Data Mining: Federal Efforts Cover a Wide Range of Uses*. Report to the Ranking Minority Member, Subcommittee on Financial Management, the Budget and International Security, Committee on Governmental Affairs, U.S. Senate, Washington. GAO-04-548.

Wayman, J. C., Stringfield, S., & Yakimowski, M. (2004). *Software enabling school improvement thought analysis of student data.* (CRESPAR Technical Report No. 67). Baltimore: Johns Hopkins University.

Weiss. D. (2006). *Descriptive Clustering as a Method for Exploring Text Collections*. PhD thesis, Poznan University of Technology, Poznan, Poland.