Librarian
Perpustakaan Sultanah Zanariah
UTM, Skudai
Johor

Sir,

CLASSIFICATION OF THESIS AS RESTRICTED
- *KNOWLEDGE DISCOVERY FOR LARGE DATABASES IN EDUCATION INSTITUTES*
ROBAB SAADATDOOST

Please be informed that the above mentioned thesis entitled " *KNOWLEDGE DISCOVERY FOR LARGE DATABASES IN EDUCATION INSTITUTES* " be classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are:

(i) Data collected is confidential.
(ii) Data given was on trust basis and cannot be revealed to public.

Thank you.
Sincerely yours,

DR.ALEX TZE HIANG SIM
N28-403-02,
UNIVERSITI TEKNOLOGI MALAYSIA,
81310 UTM SKUDAI JOHOR
+607-5532406

Note: This letter should be written by the supervisor, addressed to PSZ and a copy attached to the thesis

KNOWLEDGE DISCOVERY FOR LARGE DATABASES IN EDUCATION INSTITUTES

ROBAB SAADATDOOST

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

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

JUNE 2011

This dissertation is dedicated to my Parents who have never failed to give me every support, and my supervisor for giving all my need during the time we researched and for teaching me that even the largest task can be accomplished if it is done one step at a time.

ACKNOWLEDGEMENT

Praises to God for giving me the patience, strength and determination to go through and complete my study. I would like to express my deep and sincere gratitude to my supervisor; Dr.Alex Tze Hiang Sim, Her wide knowledge and her logical way of thinking have been of great value for me. Her understanding, encouraging and personal guidance have provided a good basis for the present thesis. I would like to express my appreciation to Assoc. Prof. Dr. Ali Selamat, Dr. Mohd Shahizan Othman, Dr. Roliana Ibrahim and Dr. Ab. Razak Che Hussin, for their support and guidance during the course of this study and Madam Lijah Rosdi for their friendly help.

During this work I have collaborated with many colleagues for whom I have great regard, and I wish to extend my warmest thanks to all those who have helped me with my work in Institute for Research and Planning in Higher Education (IRPHE).

ABSTRACT

This project presents the patterns and relations between attributes of Iran Higher Education (Iran Higher Education) data gained from the use of data mining techniques to discover knowledge and use them in decision making system of IHE. Large dataset of IHE is difficult to analysis and display, since they are significant for decision making in IHE. This study utilized the famous data mining software, Weka and SOM to mine and visualize IHE data. In order to discover worthwhile patterns we used clustering techniques and visualized the results. The selected dataset includes data of five medical university of Tehran as a small data set and Ministry of Science - Research and Technology's universities as a larger data set. Knowledge discovery and visualization are necessary for analyzing of these datasets. Our analysis reveals some knowledge in higher education aspect related to program of study, degree in each program, learning style, study mode and other IHE attributes. This study helps to IHE to discover knowledge in a visualize way; our results can be focused more by experts in higher education field to assess and evaluate more.

ABSTRAK

Disertasi ini merupakan penyelidikan terhadap pola dan hubungan antara atribut Institusi Pengajian Tinggi Iran dengan menggunakan teknik "data mining" dalam pengumpulan data untuk eksplorasi dan pencarian pengetahuan serta menggunakannya dalam sistem membuat keputusan IHE. Dataset IHE yang besar dan sukar untuk dianalisis dan paparan, kerana ia adalah penting untuk membuat keputusan dalam IHE. Penyelidikan ini menggunakan perisian "data mining" (pelombongan data) yang terkenal iaitu Weka dan SOM untuk melombong (mine) dan visualisasi data IHE. Dalam rangka mencari pola-pola yang berharga, penyelidik menggunakan teknik pengugusan, pengkelasan dan divisualisasikan hasilnya. Dataset yang dipilih merangkumi data dari lima universiti perubatan Teheran sebagai kumpulan data kecil dan Kementerian Sains - universiti Penyelidikan dan Teknologi sebagai data yang lebih besar ditetapkan. Penemuan pengetahuan dan visualisasi adalah perlu dalam proses untuk menganalisis dataset ini. Hasil analisis menunjukkan pengetahuan dalam aspek pendidikan tinggi adalah berkaitan dengan program pengajian, ijazah dalam setiap program, kaedah pembelajaran, mod pengajian dan atribut IHE yang lain. Penyelidikan ini membantu untuk IHE dalam mencari pengetahuan dengan cara pemvisualisasian dan hasil kajian boleh difokuskan lebih oleh pakar-pakar di bidang pengajian tinggi untuk menilai dan menilai lebih.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX

CHAPTER 1

INTRODUCTION

1.1 Introduction

Nowadays each organization deals with some data about their area and during time it increases, one of these organizations that includes big volume of data is higher education institute that has always held much data about universities, students and teachers. Thus, it is possible for us to discover some worthwhile relations or patterns that can be useful for making decision. For examples, planning the future development of a university and identifying the cluster of students who required more attentions. Management faces many challenges particularly in planning and for this purpose it needs some facts extracted from data. In our rapidly changing world, every year we accumulate data and add it to our data sets so after several years we will have a massive databank, in this environment every year our data volume increases so we need some tools to analysis this data for extracting some valuable outcome from it. Data mining has many techniques that can apply and facilitate analyzing of data.

Data mining has many definitions and almost all of them point to the discovery of patterns, and the analysis of some relations between variables in data. It does not limit to collecting and managing data; it also includes analysis. In this study, we intend to use historical data as the basis of discovering hidden relations. We intend to perform data mining techniques to discover knowledge. There are some examples, such as:

- Mining of statistical data of one university to discover successful students(Venus Shokorniaz & akbari, 2008)(Venus Shokorniaz & akbari, 2008) .
- Mining on students and discovering groups of students those are available from the data and their relations (Yghini, Akbari, & Sharifi, 2008)(Yghini, Akbari, & Sharifi, 2008) .

In this project, we applied data mining techniques on data related to Iran Higher Education Institute to discover some relations and patterns that are useful in decision making system of higher education.

1.2 **Problem Background**

We have chosen this topic because of government and management of universities need to plan before an event occurrence. We face huge data and need to analysis them to reach some knowledge. For this purpose we need some techniques that data mining helps us on this way, data mining has two common techniques that are classification and clustering. In this project we study about these techniques and choose one of them in our project.

Clustering is a data mining technique that is a division of data elements into groups of similar objects without advance knowledge of the group definitions. In addition it is a tool for data analysis, which solves classification problems. In clustering, there are strong associations between members of each group and according to the type of clustering; Clustering algorithms have 4 types: exclusive, overlapping, Hierarchical, and Probabilistic. We may find some associations between different groups. Some of these associations are strong and some of them are weak. For example exclusive algorithm has weak association and overlapping has strong (Berkhin, 2006). Clustering is a discovery tool that may discover associations and patterns in data which is not previously obvious. In short: clustering attempts to find some groups of elements, based on some similarities (Ong, 2000). One of the cluster analyses is SOM (self organizing method) that is one of the most important algorithms in data visualization and exploration. Visualization transforms from the invisible to the visible (Alhoniemi, et al., 2002,2003). SOM is a particular type of neural network used in clustering. It maps high dimensional input onto two dimensional.

Classification is a data mining technique that predicts data elements' group, for example we can predict the weather of a day will be sunny, rainy or cloudy. In classification we have predefined classes that classification is a task to assign instances to these classes opposite of clustering that we don't have knowledge about group definitions. In clustering we cluster elements based on their attribute on the contrary in classifying we classify elements into groups by recognizing pattern.

Our concern in this study is finding a way to discover exciting knowledge for universities management to achieve an appropriate plan to improve society. Each society can be affected by higher education; its economic, political and scientific improvement can be resulted by advanced higher education, when we develop our programs and management part in universities so its output will be successful graduated students that can be helpful in every part of one society.

1.3 Problem Statement

The main problem is the volume of data that we have in higher education and government needs to analysis and discover fast and correct knowledge from them. "How we can discover it?" is our significant question. Our endeavour was proposing a methodology to discover knowledge that reveals some patterns and relations between data. We have huge databases about universities but most of time we cannot use them efficiently because we don't have appropriate pattern for our executive system whereas we have data to this purpose.

We encounter many challenges in higher education such as allocating budget to universities in start of new semester without delay, accepting accurate number of students for every semester, finding ways for effective teaching. In this study our intension is discover some findings that help us in this way to improve higher education decisions and policies.

1.4 Project Objectives

Objectives of this study are:

- To study for an appropriate data mining approach suitable for analysing Higher Education Institute of Iran.
- To analyse data on medical universities to discover patterns usable for managerial decision making.
- To generalise steps for discovery on larger number of universities.
- To suggest an appropriate software for the analyses of this research.

1.5 Project Scope

The study is analyse huge volume of data in institute of research and planning for higher education (IRPHE) in Iran. We choose two dataset; the smaller dataset is medical universities of Tehran and the larger one is Ministry of Science - Research and Technology's universities. We proposed a methodology that can be used in similar institute in Iran with large volume of data.

1.6 The Project Importance

Knowledge discovery is necessary for most of our plantings. Nowadays planning for higher education has significant impact on developing of one society, successful planning needs to analysis some huge and historical data that is available in higher education institutes; most of the time there is not any correct and precise analysis whereas this analysis can be helpful for managers, researchers to plan, report and discover some knowledge.

The other importance of this project is collection of data with large volume that relate to many universities during many years so it has high probability to discover some exciting knowledge. This study can be used in every university, because it faces much data about students, teachers, staff and financial resources and most of the time, this data includes information and worthwhile patterns.

The methodology that will be suggested is helpful for higher education institute in collecting data and to discover some knowledge for improving

management decisions. Furthermore it can be useful for researchers that study and research about higher education. This knowledge leads us to development in technical, scientific and economic aspects according to type of data that we will analyse.

1.7 **Summary**

This chapter discussed the overview of this study which is knowledge discovery in higher education institute with statistical data and other associated issue in data mining techniques that helps us in discovering relations and patterns that are useful in decision making in higher education. Higher education usually does not have appropriate methodology to analysis huge volume of data and thus most of time it ignores much knowledge that can be very helpful in many decisions to improve quality of education. There are four objectives that were achieved successfully in order to improve higher education decision making systems and their policies that certainly affect on society. At first we studied knowledge discovery and then analysed Iran higher education data to find some relations and patterns. Finally we proposed a methodology to knowledge discovery from higher education data.

REFERENCES

Al-Noukari, M., & Al-Hussan, W. (2008). *Using Data Mining Techniques for Predicting Future Car market Demand; DCX Case Study.* Paper presented at the Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on.

Aldana, W. A. (2000). A Brief History of Data Mining.

Alhoniemi, E., Himberg, J., Hollm´en, J., Laine, S., PasiLehtim¨aki, Raivio, K., et al. (2002,2003). *SOM in data mining.*

Becerra-Fernandez, I., Zanakis, S. H., & Walczak, S. (2002). Knowledge discovery techniques for predicting country investment risk. *Computers & Industrial Engineering, 43*(4), 787-800.

Bellinger, G., Castro, D., & Mills, A. (2010). Data, Information, Knowledge, and Wisdom.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques.

Cios, k. j., pedrycz, w., swiniarski, r., & kurgan , l. A. (2007). Data Mining- A knowledge discovery approach.

Delavari, N., & Beikzadeh, M. R. (2004). A new model for using data mining in higher educational system.

Delavari, N., Beikzadeh, M. R., & Amnuaisuk, S. (2005). Application of enhanced analysis model for data mining processes in higher educational system.

Famili, F. (2009, 27-28 Oct. 2009). *Knowledge discovery and management in life sciences: Impacts and challenges.* Paper presented at the Data Mining and Optimization, 2009. DMO '09. 2nd Conference on.

Fangjun, W. (2010, 6-7 March 2010). *Apply Data Mining to Students' Choosing Teachers Under Complete Credit Hour.* Paper presented at the Education

Technology and Computer Science (ETCS), 2010 Second International Workshop on.

Fayyad, U. M., & Uthurusamy, R. (1996). Data mining and knowledge discovery in databases.

Ghanbari, M. (2007, 4-6 March 2007). *Visualization Overview.* Paper presented at the System Theory, 2007. SSST '07. Thirty-Ninth Southeastern Symposium on.

Guruler, H., Istanbullu, A., & Karahasan, M. (2010). A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education, 55*(1), 247-254.

Hilderman, R. J., Liangchun, L., & Hamilton, H. J. (1997, 3-8 Nov 1997). *Data visualization in the DB-Discover system.* Paper presented at the Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on.

Himberg, J. (1999). SOM based cluster visualization and its application for false coloring.

Holmes, G., Donkin, A., & Witten, I. H. (1994). Weka: A machine learning workbench.

Jakulin, A. (2010).

Koua, E. L. (2005). USING SELF-ORGANIZING MAPS FOR INFORMATION VISUALIZATION AND KNOWLEDGE DISCOVERY IN

COMPLEX GEOSPATIAL DATASETS. *Seminar on Data and Information Management*

SS 2005 2D, 3D and High-dimensional Data and Information Visualization.

Lam, S. B. (2007). DATA MINING WITH CLUSTERING AND CLASSIFICATION.

LAROSE, D. T. (2005). DISCOVERING KNOWLEDGE IN DATA.

Mannila, H. (1996, 18-20 Jun 1996). *Data mining: machine learning, statistics, and databases.* Paper presented at the Scientific and Statistical Database Systems, 1996. Proceedings., Eighth International Conference on.

Mierle, K., K., L., Roweis, S., & Wilson, G. (2005). Mining student CVS repositories for performance indicators.

Mitchell, T. M. (1997). Machine learning.

Ong, C. S. (2000). KNOWLEDGE DISCOVERY IN DATABASES: AN INFORMATION RETRIEVAL PERSPECTIVE.

Pazzani, M. J. (2000). Knowledge discovery from data? *Intelligent Systems and their Applications, IEEE, 15*(2), 10-12.

Salazar, A., Gosalbez, J., Bosch, I., Miralles, R., & Vergara, L. (2004, 28 June-1 July 2004). *A case study of knowledge discovery on academic achievement, student desertion and student retention.* Paper presented at the Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on.

Saraee, M. H., & Theodoulidis, B. (1995, 1 Feb 1995). *Knowledge discovery in temporal databases.* Paper presented at the Knowledge Discovery in Databases, IEE Colloquium on (Digest No. 1995/021 (A)).

Siraj, F., & Abdoulha, M. A. (2009). *Uncovering Hidden Information Within University's Student Enrollment Data Using Data Mining.* Paper presented at the Modelling & Simulation, 2009. AMS '09. Third Asia International Conference on.

Tan, Steinbach, & Kumar. (2006). Introduction to Data Mining.

Tsai, P. S. M., & Chen, C.-M. (2001). Discovering knowledge from large databases using prestored information. *Information Systems, 26*(1), 1-14.

Vembu, S., & Baumann, S. (2004). A Self-Organizing Map Based Knowledge Discovery for Music Recommendation Systems. *Computer Music Modeling and Retrieval Second International Symposium, CMMR 2004, Esbjerg, Denmark, May 26-29, 2004, 3310*, 119-129.

Venus Shokorniaz, & akbari, A. H. a. (2008). Mining of statistical data of one university to discover successful students.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the Self-Organizing Map. *IEEE TRANSACTIONS ON NEURAL NETWORKS, 11.*

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (1999). Self-organizing map in Matlab: the SOM Toolbox. *Proceedings of the Matlab DSP Conference 1999, Espoo, Finland, November 16–17, pp. 35–40, 1999.*

Witten, I. H., & Frank, E. (2005a). Data Mining :Practical Machine Learning Tools and Techniques,.

Witten, I. H., & Frank, E. (2005b). Data Mining: Practical machine learning tools and techniques, 2nd Edition.

Yghini, M., Akbari, A., & Sharifi, M. (2008). Mining on students and discovering groups of students those are available from the data and their relations.

Zanakis, S. H., & Becerra-Fernandez, I. (2005). Competitiveness of nations: A knowledge discovery examination. *European Journal of Operational Research, 166*(1), 185-211.