J. Basic. Appl. Sci. Res., 4(6)28-37, 2014

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Knowledge Discovery in Higher Education Institutes

Robab Saadatdoost[1,2,*], Alex Tze Hiang Sim[1], Jee Mei Hee[4], Hosein Jafarkarimi[1,3], Leila Saadatdoost[5]

[1]Department of Information Systems, Faculty of Computing, Universiti Teknologi Malaysia
[2]Department of Computer and Information Technology, Parand Branch, Islamic Azad University, Parand, Iran
[3]Department of Computer, Damavand Branch, Islamic Azad University, Damavand, Iran
[4]Department of Educational Foundation, Faculty of Education, Universiti Teknologi Malaysia
[5]Department of Computer engineering, Sari Branch, Islamic Azad University, Sari, Iran

Received: March 4, 2014
Accepted: May 8, 2014

ABSTRACT

This paper documents our endeavor in utilizing clustering techniques based on 18 years data gathering in Persian to discover interesting findings from five medical universities in Tehran, Iran. Research reveals findings in related to the opportunities and concerns in students' (I) perceived acceptances on study programs, (II) preferred study mode, and (III) programs that could be offered to perspective students. The discovered knowledge could be utilized to improve ranking of a university, in specific, and demonstrated the application of data mining in education.
KEYWORDS: knowledge discovery, data mining, higher education institutions, clustering, visualization.

INTRODUCTION

Higher Education Institutions (HEI) plays a vital role in supporting a country's development by educating the next generations. Each year, higher education institutions produce a large number of graduates and store a large number of records concerning the students and programs offered.

Typically, there is much effort in collecting data by HEI. For example, students are evaluated and their marks for all courses in a program are calculated, reported and stored. With the advancement of technology, data is valuable and should no longer be kept for archiving purposes alone. These data, if carefully processed, offer opportunities for the advancement of a higher education institution. However, there is little research[1] in analyzing historical data for strategic and management values.

The process to understand the value of data via this state of the art technology is termed as knowledge discovery. It is a process to identify hidden patterns in a data set which are useful for decision making and predictions[2]. Indeed, knowledge discovery is a computer science concept which describes the searching process of large volumes of data for knowledge or patterns about the data[3]. Besides, knowledge discovery is an essential means to detect the valid and useful knowledge[4]. Interesting knowledge can be discovered although these knowledge are typically hidden from general and typical analysis [5][6]. During the discovery process, a number of data mining techniques are utilized aiming to discover patterns and relations from raw data. The discovered knowledge often offers new perspectives from databases [7].

In this research, we applied data mining techniques on data of universities. The data was accumulated from 1988 to 2005 on five larger medical universities of Iran, Middle East. The results of data mining revealed facts about the country's education policies and development in general. Among these results are the differences in distributions of programs among universities, detection of hidden opportunities and some concerns for the universities to improve its ranking. Based on the information, these findings are critically important for the top management to maintain the existing management decision and to plan for new discourse.

LITERATURE REVIEW (DATA MINING IN EDUCATION)

There are limited research findings on discovering knowledge from HEI data. Potawat et al. [1] analyzed student profile to understand about students enrolment. To improve management effectiveness, Delavari and Beikzadeh[8] proposed a model and detailed guidelines in using data mining in higher education system. Delavari designed a roadmap draft for the application of data mining in higher education system [9] and based on the roadmap, HEI decision makers are enabled to detect and enhance the existing processes using appropriate data mining techniques and procedures. In addition, Mugla University, Turkey has developed an advanced system

***Corresponding Author:** Robab Saadatdoost, PhD Information Systems Candidate, Department of Information Systems, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 UTM Johor, Malaysia, Email: saadatdoost@gmail.com

named MUSKUP (Mugla University Student Knowledge Discovery Unit Program) to analyze the available data that stored in their conventional database management system [10]. This approach provides access to all functions through a single software. Based on their study, hidden relations were found among different types of registration, groups of family economic and their success in studies.

In another case, Mierle et al. [11] proposed a model and is used to discover students' behavior during the encoding in the computer programs. The discovered patterns were used to enhance student's ability in writing the codes. In another research, authors [12] used knowledge discovery about academic achievement, student retention and desertion to generate a set of decisions. For examples:

(1) Majority of students that enter university with HIGHER AGE may drop out at ANY level of study.
(2) Students that enter the university with LOWER AGE have high probabilities to gain GOOD academic performance.

Based on the decision rules, they were able to propose strategies to improve the existing academic processes[12].

Similar research based on enrolment data of Sebha University, the authors Libya, Siraj and Abdoulha[13] discovered knowledge that could be used as a guideline to detect, improve and streamline the business processes. For example, the reasons for low enrolment were detected across several branches of the university. Possible correlations among attributes such as, faculty, gender and status of the enrolled student (e.g. active, graduated, and transferred to another program) were also identified.

In 2000 , a success research was done by School of Information Technology, Jiangxi University, China. Their study showed students choose supervisors based on four different criteria: knowledge-seekers, rational, emotional and having the sense of indifference [14]. Based on the finding of the study, the school could regroup the present students and identify effective factors in choosing appropriately supervisors.

Recently, a research was done proposing the use of knowledge discovery to extract knowledge by analyzing data provided by the interaction of students with e-learning environments such as Moodle. Teachers sometimes become overwhelmed and unable to process the data without the support of techniques and tools that are useful for analyzing large data flows. Their proposal creates historical reference models of students that completed the course or dropped out of it. These models can be used to classify a specific student within the dropout or non-dropout group[15].

In another case, [16] studied factors affecting english language learning in university students by data mining. They found a model to improve and increase the university students' success in English language. They used classification and clustering algorithms are used to analyze data by Data Mining[16].

In 2014, I. Bhaduri and S. Bhaduri used knowledge discovery in school databases to enrich learning process, the large databases of learners needs to be worked on to discover the evolving patterns of learners and to enhance the course of learning.The purpose of research was to use data mining to create a valid student assessment [17].

Data mining in education is relatively a scarce topic, especially in collecting good data and dealing with unfamiliar topic. In the next section, we explain an approach for mining data in HEI.

METHODOLOGY

Data mining research are mostly based on an industrial standard methodology termed as Cross Industry Standard Process for Data Mining (CRISP–DM) [18] which consists of mainly six phases – business understanding, data understanding, data preparation, modeling, evaluation and deployment. Much time, even up to 80% of the entire project, is spent on activities before modeling the problems and its solutions. As shown in the Figure 1, our research methodology (modified from CRISP-DM) consists of two parts with six phases.

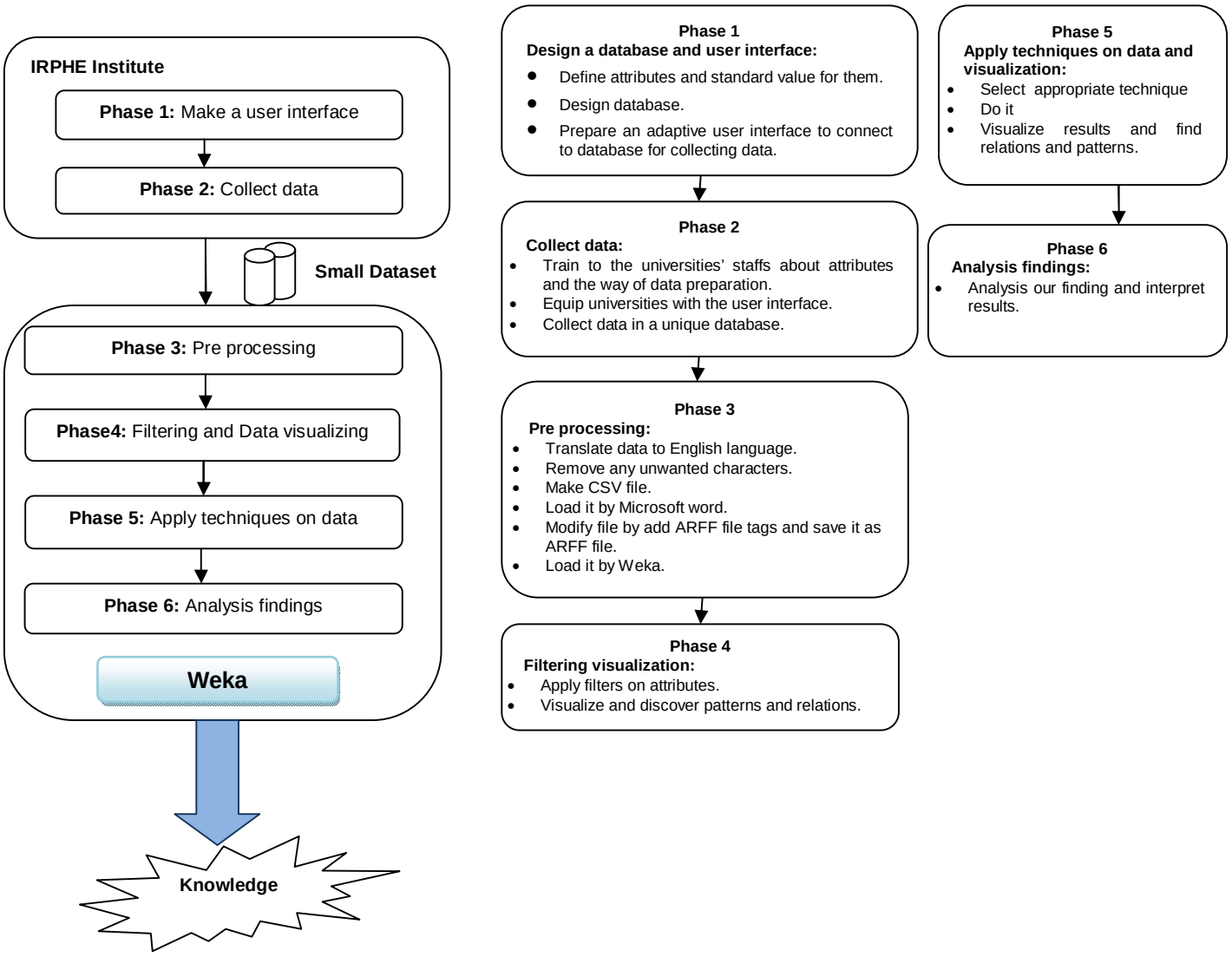

Figure 1. : Methodology

In the first part of the study, five medical universities data were previously collected via the Institute of Research and Planning for Higher Education (IRPHE), which is affiliated with the Ministry of Science, Research and Technology, Iran. IRPHE has well understood all business requirements and collected data that are important. They have, in fact, created a DBMS and conducted training to HEI in order to obtain and store quality data for the past 18 years.) The researchers do not have the budget and time to fine tune or collect data; however, we have a close collaboration with them which benefits both parties. Our primary focus is to undercover knowledge from established and reliable sources. In this case, we begin from understanding the data and further prepare the data for data mining. We give emphasis on two unique activities in the second part of our methodology:

(i) During the pre-processing phase, we have utilized Google Translator (http://translate.google.com) to translate Persian into English.

(ii) We have carefully selected Weka, an open source machine learning tool to filter and visualize our data based on K-means algorithm. We will later compare our results with another visualization algorithm termed as Self-organizing map (SOM)

Dataset

Our data covers a group of five medical universities from year 1989 to 2006. These are listed below:

1. Tehran University of Medical Sciences
2. Iran University of Medical Sciences
3. Shaheed Beheshti University of Medical Sciences

4. Institute of Pastor Iran
5. Blood Transfusion Institute

The dataset contains 3,813 records with the following 10 attributes: (i) Year, (ii) Name of university, (iii) Program, (iv) Degree, (v) Learning style, (vi) Type, (vii) Study mode, (viii) Number of woman, (ix) Number of man, and (x) Number of students.

MODELING

The K-Means algorithm requires a pre-determination on the number of knowledge clusters that could be discovered. For example, groups for universities that is to be considered similar to one another. We have tried various values for K and each has the potential to discover different patterns. To provide a standard, according to experts' experience, we utilized EM algorithm to suggest the best number of clusters after cross-validating the values automatically [19]; EM algorithms generated K=13 in our research (for an effective division of clusters). This maximized the knowledge that could be discovered from dataset.

Using the higher K values (compare to our intuitive values ranging between 2 - 5), universities like Tehran, Iran and Shaheed Beheshti University of Medical Sciences were divided into smaller subgroups based on the 10 attributes in section dataset. In the following sections, we show the various models based on this data set with its analyses.

ANALYSIS AND RESULTS

We discovered a number of interesting findings listed below. It is interesting to note that although we are not the management of any university, it is as though we have interviewed the management from different universities or becoming experienced management personnel to report on the following 'stories'. This is made possible through careful formulation of methodology which includes selection of data mining techniques and parameter settings.

(1) Surprisingly, a number of the new(er) and perceived to be important study programs such as Children Dental, Orthodontics, Nerve Disease, General Surgery were terminated for many years ever since year 1993. Termination of programs involved cost and time lost, resigning personnel and the issue of image. Further investigation suggests that a possible cause could be due to the introduction of Pathology programs. (This is of course unknown to the management when Pathology programs were first introduced.) Other smaller universities should learn from this experience so as not to offer programs that coincide with Pathology in view of limited capital. In short, the Pathology programs are preferred over other programs listed above in the five larger universities. See Figure 2

Figure 2.Program as X axis – Year as Y axis. The distribution depicts opportunities to offer a certain programs.

On the opportunities:

(2) Opportunity (a): The two larger universities (Iran University of Medical Sciences and Shaheed Beheshti University of Medical Sciences) could offer programs which include Clinical Pharmacy, Thoracic Surgery, Radiology-Oral, Maxillofacial, and Oral Diseases. These programs could have been offered to prospective students to tap on the high demand. See Figure 3

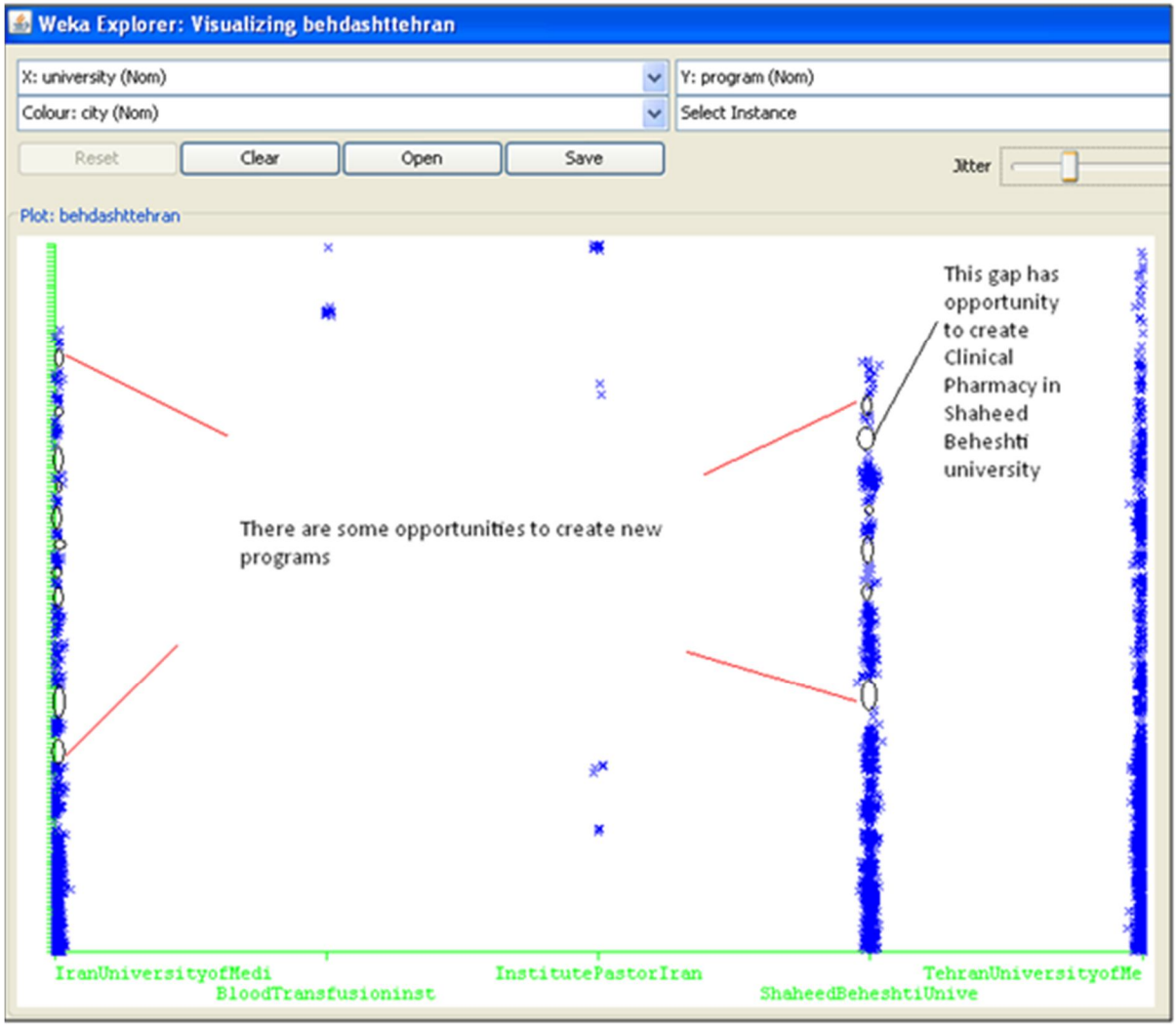

Figure 3.University as X axis – program as Y axis. The gaps show opportunities to offer certain programs in each university.

(3) Opportunity (b): Rather hidden among the data, we discovered an emerging group of students. With the current trend, more and more students are taking 'overnight study' (paid evening classes). This trend of study influences the way of running courses in universities, and should be considered by management to tap on its benefits. See Figure 4

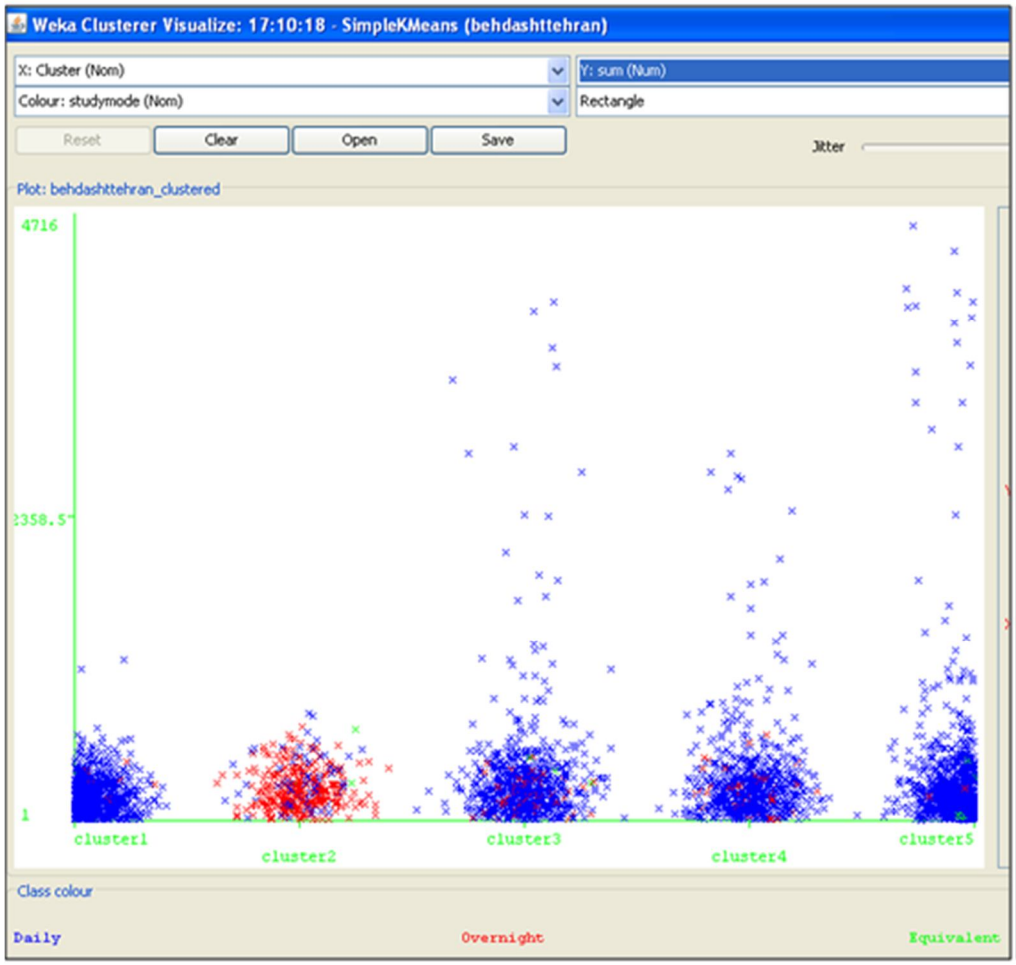

Figure 4. Cluster as X axis – Sum of students as Y axis– Study Mode as Color

There are concerns as follows:

(4) Concern (a): A large number of programs are taught only to the postgraduates but its fundamentals are not available to the undergraduates among the five universities. This suggested that postgraduate students are originated from the other smaller universities. This scenario should be reassessed if the undergrads are to be taught by the same universities meeting the students' need and to gain popularity in the field. (Another argument is these universities only teach postgraduates due to limited manpower. If this is the case, students from smaller universities should learn from this experience in designing their programs.) See Figure 5

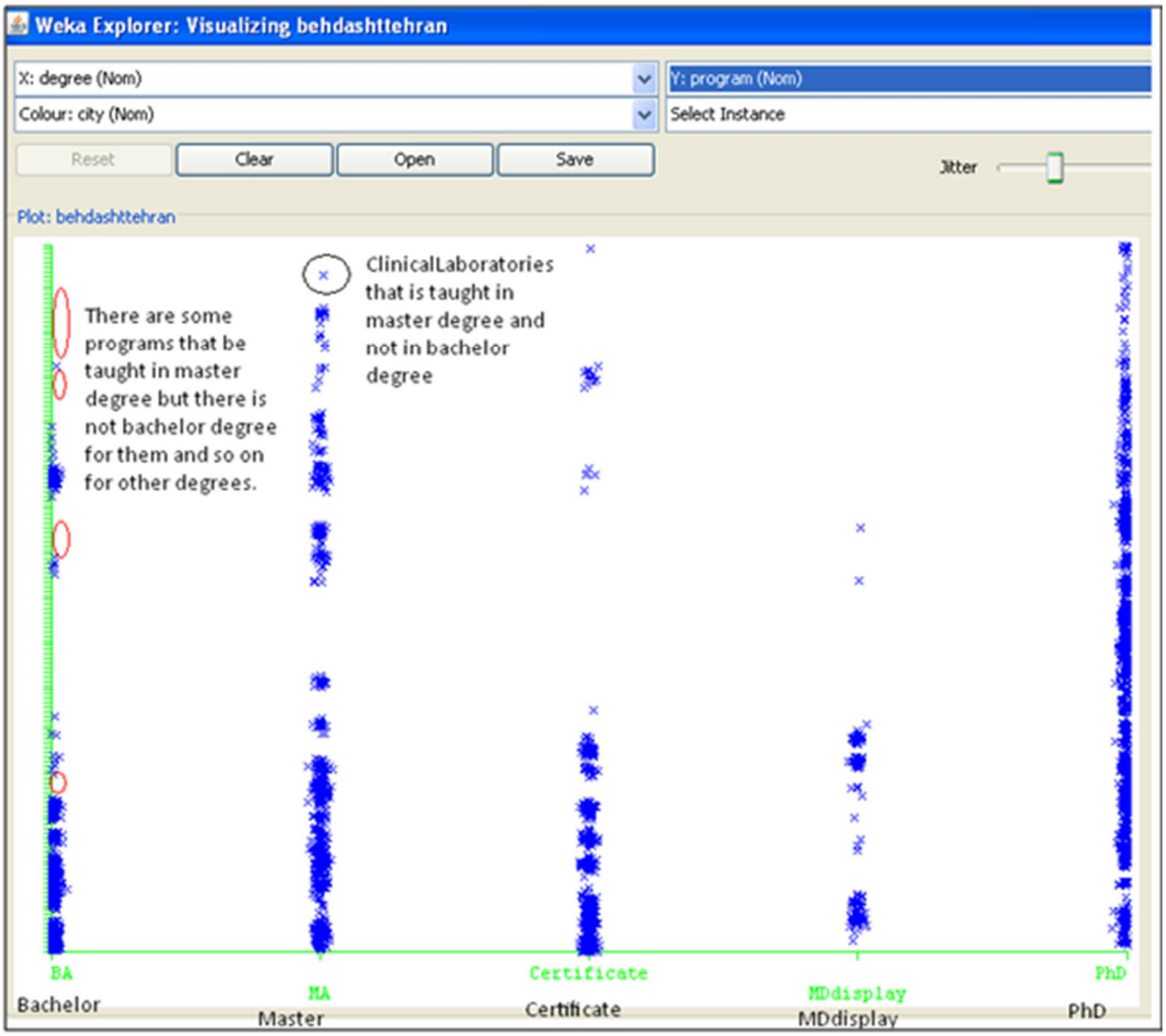

Figure 5.Degree as X axis – Program as Y axis. A comparison of programs among bachelor, master, certificate, MD display and PhD degree would reveal the discontinuity of some programs.

(5) Concern (b): There is a huge demand for the program Doctor of Medicine (named as 'MD-display') compared to 'Master Degree in Medical'. The former is a postgraduate program being independent from the latter and it is designed for selected students to continue their bachelor degree. (We also believe the prospect of being a doctor is good for younger generations.) The latter is a typical standalone program for adult learners. In view of this, perhaps more resources could be directed to the former program under the constraints of budget. See Figure 6

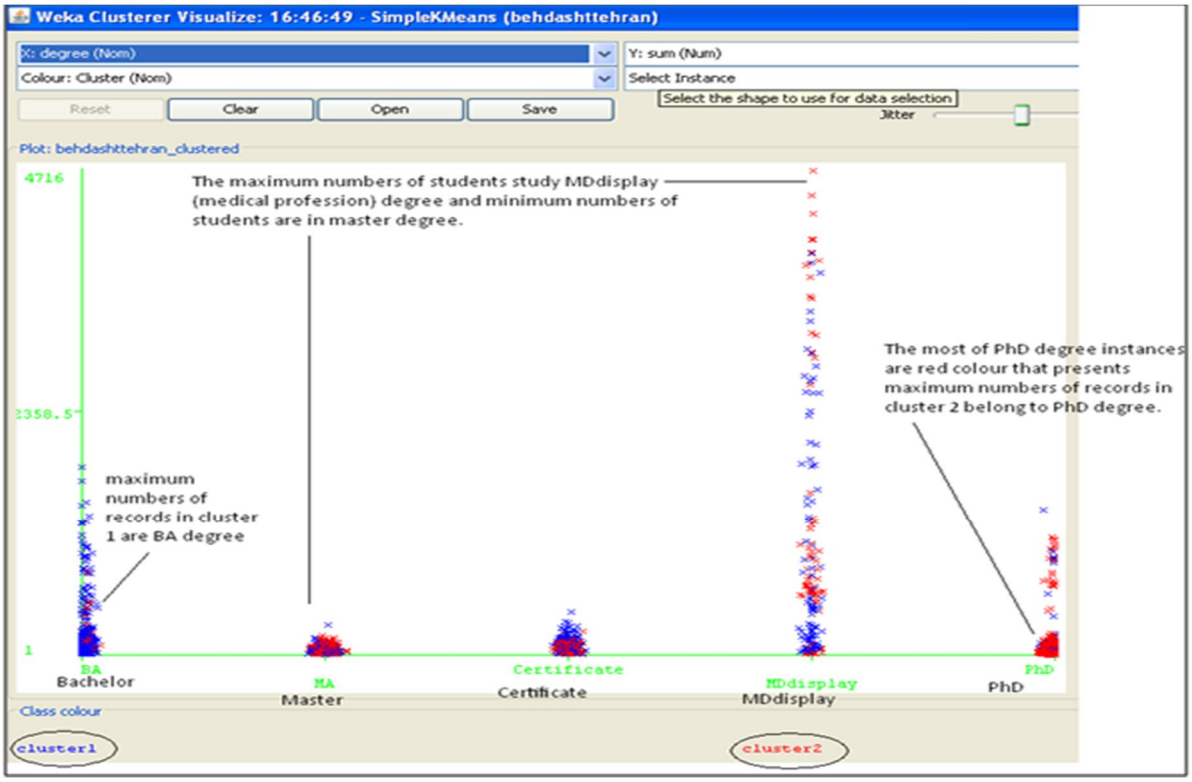

Figure 6. Degree as X axis– Sum of students as Y axis

(6) Concern (c): The new 'overnight study' mode is mainly provided by three universities. Many students seem to perceive a much loose entrance requirement to study in this mode compare to a typical 'daily mode'. This has contributed to a large number of students in one of the major University of Medical Sciences. Such perception (regardless of its validity) may not be good for the advancement of the university and should be studied by the management accordingly. See Figure 7

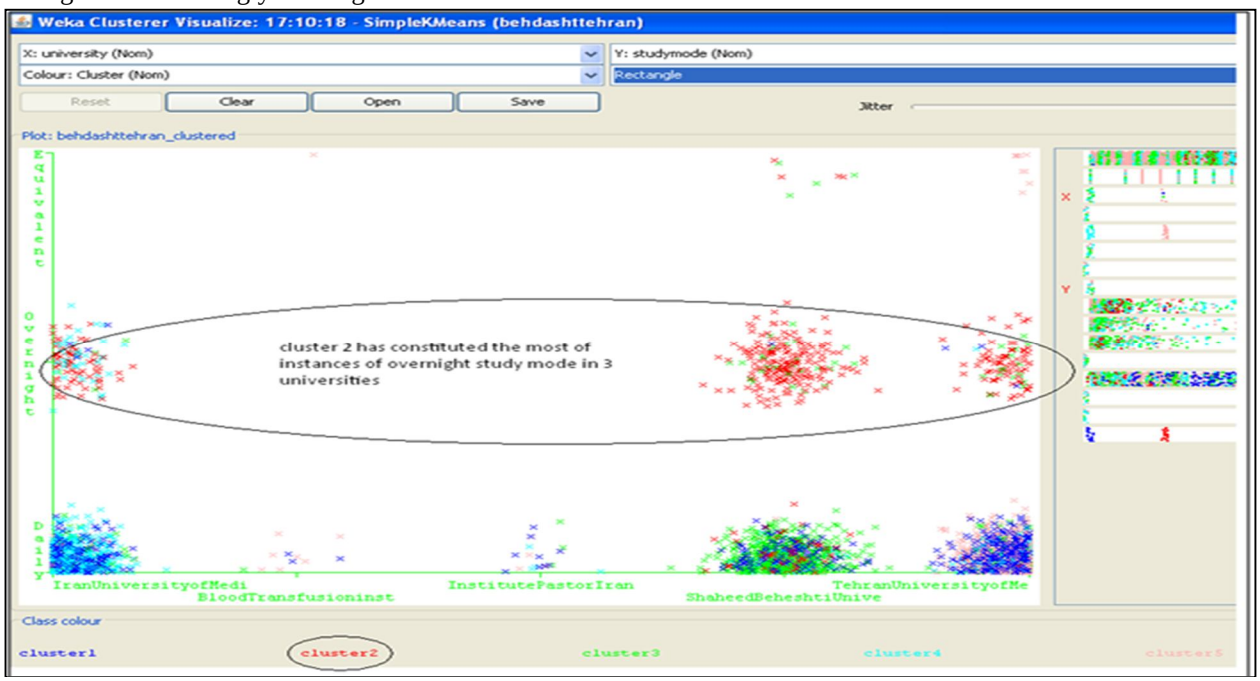

Figure 7.University as X axis – Study type as Y axis. Cluster 2 consists of a large number of 'overnight' study mode as offered by three universities.

A COMPARISON TO SOM RESULTS

SOM being an unsupervised neural network technique also known as Kohonen's self-organizing map decreases high dimensional vectors into lower dimensional maps. (The steps to use SOM can be referred to[20][21])

We show our SOM map in Figure 8- 9. We conclude that K-Means could reveal more knowledge compare to SOM but SOM could show an emerging trend easily. For example, an observation from the rectangle in the 'study mode' plane with its density colored red, and 'year' plane shows the 'overnight' mode of study has appeared recently. The overnight mode of study is a paid program with certain perceptions (see point no. 6, Section analysis and result), and it becomes more popular, technically this strengthens our K-Means discovery.

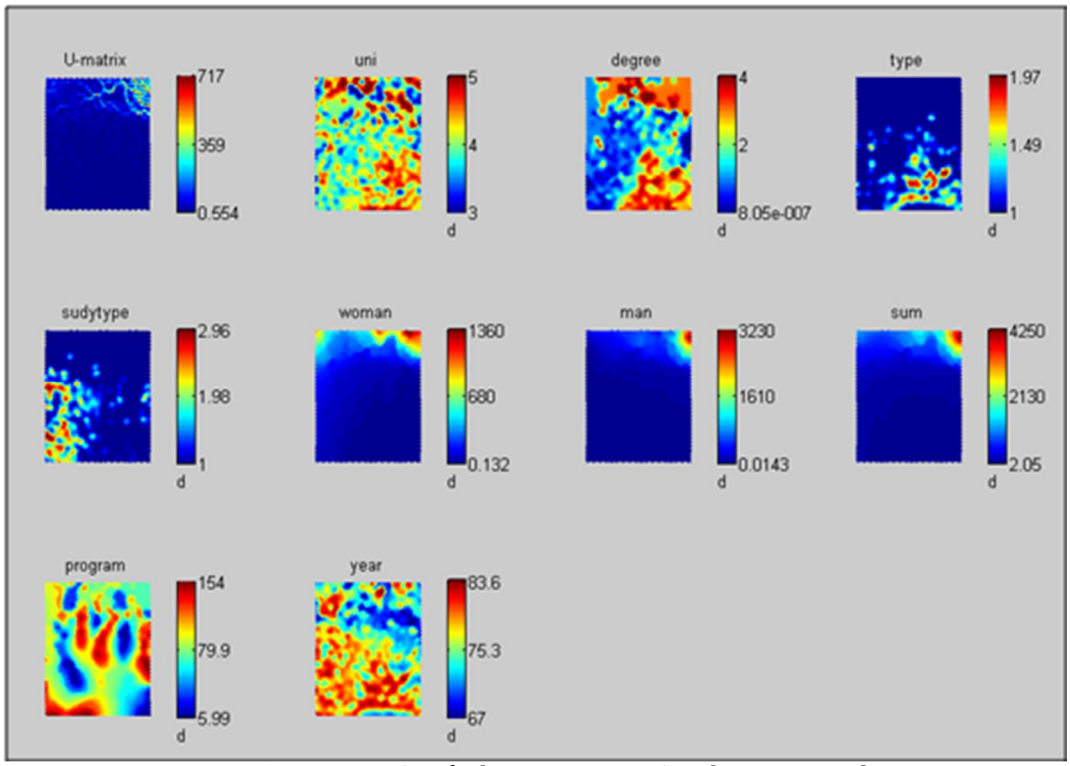

Figure 8. U-Matrix (Unified Distance Matrix) and component planes

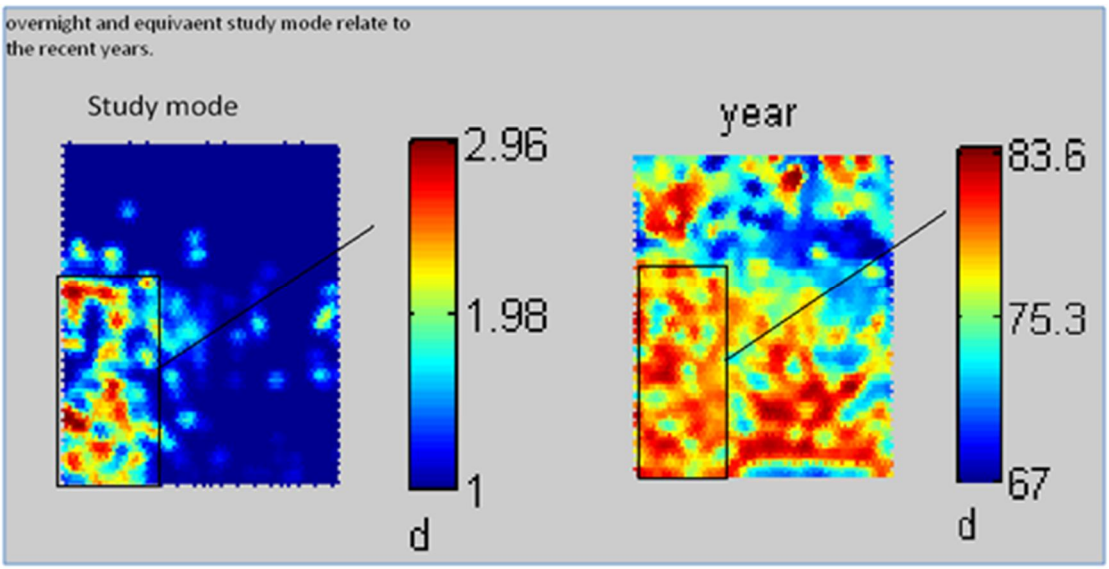

Figure 9. Study mode and year component planes

CONCLUSION AND FUTURE WORK

We contribute to the field of data mining in the management of HEI in view of a scarce literature in the field. We have demonstrated a sound methodology to discover knowledge and detailed out the opportunities, concerns and trends that could be discovered by the top management. Personally, we answered our curiosity on 'what are the samples of important knowledge that could be discovered from data kept by HEI for all this while?' With various findings on the preferred study programs, perceived acceptances and study mode, we conclude that this research project is helpful in providing insights that needed by the HEI. We are adopting the same methodology over other smaller universities to gain better understanding on the education of Iran.

ACKNOWLEDGMENT

The author A.T.H. Sim would like to acknowledge in received of Research University Grant (vot. 07J79) from Ministry of Higher Education (MOHE) and Universiti Teknologi Malaysia (UTM).

REFERENCES

1. Potawat, V., et al., Using rough set theory for automatic data analysis. 29th congress on science and technology of Thailand., 2003.
2. Chen, C.-H., et al., Knowledge discovery using genetic algorithm for maritime situational awareness. Expert Systems with Applications, 2014. 41(6): p. 2742-2753.
3. Tan, H., Knowledge Discovery and Data Mining. Vol. 135. 2012: Springer.
4. Cruz, J., et al. Knowledge discovery in virtual worlds usage data: approaching web mining concepts to 3d virtual environments. in Proceedings Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support (Eureka-2013). 2013.
5. Famili, F. Knowledge discovery and management in life sciences: Impacts and challenges. in DMO '09. 2nd Conference on Data Mining and Optimization. 2009.
6. Sangster-Gormley, E., et al., Use of Knowledge Discovery Techniques to Understand Nurse Practitioner Practice Patterns and Their Integration into a Healthcare System. Enabling Health and Healthcare Through ICT: Available, Tailored and Closer, 2013. 183: p. 111.
7. Al-Noukari, M. and W. Al-Hussan. Using Data Mining Techniques for Predicting Future Car market Demand; DCX Case Study.in Information and Communication Technologies: From Theory to Applications. 2008.
8. Delavari, N., M.R.A. Shirazi, and M.R. Beikzadeh. A new model for using data mining technology in higher educational systems.in Information Technology Based Higher Education and Training. ITHET . 2004.
9. Delavari, N., M.R. Beikzadeh, and S. Amnuaisuk, Application of enhanced analysis model for data mining processes in higher educational system. 2005.
10. Guruler, H., A. Istanbullu, and M. Karahasan, A new student performance analysing system using knowledge discovery in higher educational databases. Computers & Education, 2010. 55(1): p. 247-254.
11. Mierle, K., et al., Mining student CVS repositories for performance indicators. 2005.
12. Salazar, A., et al. A case study of knowledge discovery on academic achievement, student desertion and student retention.in Information Technology: Research and Education. ITRE. 2004.
13. Siraj, F. and M.A. Abdoulha. Uncovering Hidden Information Within University's Student Enrollment Data Using Data Mining. in Third Asia International Conference on Modelling & Simulation. 2009.
14. Fangjun, W. Apply Data Mining to Students' Choosing Teachers Under Complete Credit Hour. in Second International Workshop on Education Technology and Computer Science (ETCS). 2010.
15. Lara, J.A., et al., A system for knowledge discovery in e-learning environments within the European Higher Education Area–Application to student data from Open University of Madrid, UDIMA. Computers & Education, 2014. 72: p. 23-36.
16. Naghdeforoshha, M. and F. Taherkhani, Studying Factors Affecting English Language Learning in University Students by Data Mining. Journal of Basic and Applied Scientific Research, 2013.
17. Bhaduri, I. and S. Bhaduri, Use of Knowledge Discovery in School Databases to Enrich The Learning Process. INTED2014 Proceedings, 2014: p. 3893-3899.
18. Shearer, C., The CRISP-DM Model: The New Blueprint for Data Mining. J.Data Warehousing, 2000.5(4).
19. Witten, I.H. and E. Frank, Data Mining :Practical Machine Learning Tools and Techniques,. 2005.
20. Alvarez-Guerra, E., et al., A SOM-based methodology for classifying air quality monitoring stations. Environmental Progress & Sustainable Energy, 2011.
21. Vesanto, J., et al., Self-organizing map in Matlab: the SOM Toolbox, in Matlab DSP Conference ,Espoo,Filand. 1999. p. 35-40.