Pattern Discovery Using K-Means Algorithm

Almahdi Mohammed Ahmed
Department of Computer Science, Faculty of Science
Sebha University Libya
Libya
sheriftsm@gmail.com

Norita Md Norwawi
Faculty of Science and Technology
Universiti Sains Islam Malaysia
Nilai, Ng. Sembilan, Malaysia
nmn@uum.edu.my

Wan Hussain Wan Ishak
School of Computing
Universiti Utara Malaysia
Sintok, Kedah, Malaysia
hussain@uum.edu.my

Ahmed Alkilany
Department of Computer Science, Faculty of Science
Sebha University Libya
Libya
aassaid@gmail.com

***Abstract*—Student's placement in industry for the industrial training is difficult due to the large number of students and organizations involved. Further the matching process is complex due to the various criteria set by the organization and students. This paper will discuss the results of a pattern extraction process using a clustering algorithm that is k-means. The data use consists of Bachelor of Information Technology and Bachelor in Multimedia students of Universiti Utara Malaysia from the year 2004 till 2005. The experiments were conducted using undirected data and directed data. The pattern extracted gave information on the previous matching process done by the university.**

Keywords-Pattern discovery;Data Mining; Clustering; k-means

I. INTRODUCTION

Industrial placement or practicum is a program where final year student undergoes a practical training in actual working environment for a certain period of time. During that period students will do the task as an actual employee of a company. They hold a certain responsibility for the company and at the same time fulfilling their program requirement [1].

Student placement to the appropriate organization is one of the crucial but tedious tasks [2]. Depending on the policy of the university, some university encourages students to find their own placement, while some other institutions arrange the placement on behalf of the students. At Universiti Utara Malaysia (UUM) students are currently encouraged to find their own placement. However in a case where the students fail to find their own placement the university will find the appropriate organization for the student.

The placement process involves a matching between student profile and organization requirements [1]. Through this process, a list of students that are fulfilling the organization requirements will be proposed. Typically, the matching process is based on certain criteria in order to serve best the organization and student. For example, a student who lives in Kuala Lumpur should not be sent to an organization located in Alor Star. This is to avoid problems in terms of accommodation, financial, and social.

The matching procedures for students and organizations involve several steps. First, the registered city1 (is the first choice for students) and city2 (is the second choice for students) will be examined. A match between organization location and student's hometown will be determined. The next criterion is the student majoring and cumulative grade point average (CGPA). Usually, the organization will request student with a specific majoring that suit with their needs and holding good CGPA. Additionally, due to certain work prospect, some organization request student based on certain gender and race. These criteria have to be considered by the program coordinator during the placement process to ensure the right student being sent to the right organization. In order to make the process easier, the program coordinator can refer to the previous placement that stored in the practicum database. After years of implementation of practicum program, the university practicum database has grown to a large and complex info structure. Thus, findings, retrieving and connecting the patterns for the current placement has become quite tedious and time consuming.

This paper discusses the application of data mining technique to extract patterns from the practicum database. K-means algorithm is used to cluster the patterns into several clusters. This pattern will be a useful guideline for future organization and student matching. The next section will discuss related literature. Following is the methodology and experiment and findings in the next section.

II. LITERATURE REVIEW

One of data mining techniques is clustering which is widely used in pattern discovery. K-means is one of the clustering algorithms that have been widely studied. Clustering refers to techniques for grouping similar objects in clusters. Formally, given a set of dimensional points and a function A * B = C that gives the distance between two points in the B, required to compute cluster centers, such that the points falling in the same cluster are similar and points that are in different cluster are dissimilar. Most of the initial clustering techniques were developed by statistics or pattern recognition communities, where the goal was to cluster a modest number of data instances. Developing clustering algorithms to effectively and efficiently cluster rapidly growing datasets has been identified as one of important challenges [3].

Lai et al [4] proposed fast k-means clustering algorithm using the displacements of cluster centers to reject unlikely candidates for a data point. The computing time of the proposed algorithm increases linearly with the data dimension d, whereas the computational complexity of major available kd-tree based algorithms increases exponentially with the value of d. Theoretical analysis shows that the method can reduce the computational complexity of full search by a factor of SF and SF is independent of vector dimension. The experimental results shows that compared to full search, the proposed method can reduce computational complexity by a factor of 1.37–4.39 using the data set from six real images. Compared with the filtering algorithm, which is among the available best algorithms of k-means clustering.

K-means has been implemented in various applications such as industrial application. Mart et al [5] discussed two clustering techniques, k-means and Fuzzy C-means, for the analysis of the electricity prices time series. Both algorithms focused on extracting useful information from the data with the aim of model the time series behavior and find patterns to improve the price forecasting. The main objective of their study was to find a representation that preserves the original information and de-scribes the shape of the time series data as accurately as possible. This work demonstrates that the application of clustering techniques is effective in order to distinguish the day groups. The groups are (1) includes the working days and (2) includes weekends and festivities.

III. Methodology

Data for the project is obtained from UUM practicum database consisting of all information technology (IT) and multimedia undergraduate students from the year 2004 till 2005. The data is prepared based on Knowledge Discovery in Databases (KDD) methodology. KDD consists of several steps: Selection, Pre-processing, Transformation, Pattern Extraction using data mining, and Interpretation/Evaluation [6].

The initial data contain the performance profile gathered from a number of 998 students with 20 listed attributes which include Metric Number, program, Session, Major, Program Code, City1, City2, Ad-dress, Address State, CGPA, Gender, Race Code, Race, Organization, Address1, Address2, Address3, Address4, Postcode, City3 and State. The data contains various types of values either string or numeric value. The target is represented as Organization's name. The Organization's name was grouped according to three categories (Government, Private, and Government_owned). Based on the discussion with the program coordinator, all 998 data are used in this study. The selected attributes Majoring, CGPA, Gender, City1, Race, Organization and City3 are chosen based on the suitability of the condition and problems presented in this paper.

The data is then pre-processed to remove the outlier and missing values. Later, some of the attributes are transformed into classes. The transformation was applied to ensure that the data mining process can be easily performed besides ensuring a meaningful result produced. The following rules are used to transform the CGPA to string data.

1. If the CGPA = 2.0 Till 2.49 THEN Replace CGPA by CLASS4
2. If the CGPA = 2.5 Till 2.99 THEN Replace CGPA by CLASS3
3. If the CGPA = 3.0 Till 3.49 THEN Replace CGPA by CLASS2
4. If the CGPA = 3.5 Till 4.00 THEN Replace CGPA by CLASS1

The transformation has also been applied to attributes city1 and city3 by grouping several cities together according to their location or region, decoded into new region using code of each state. For example, ALOR SETAR and JITRA have the same code 02 then they were converted into one Region (N_Region). The organization's name was also transformed by into three categories (Government, Private and Government_owned).

After all pre-processing and transformation have been implemented, the data were then ready to be clusters using k-means. K-means divides data into groups (cluster) that are meaningful, useful, or both. Since the meaningful group is the goal, then the cluster should capture the natural structure of the data. The k-means clustering technique has initial centroids, where k is a user specified parameter, namely the number of clusters desired. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster [7].

IV. Experiment and Findings

The experiment was conducted using k-means clustering algorithm. The aim of the experiment is grouped (cluster) the data. The sizes of clusters are set and chosen randomly from 5, 10, 13, 15, 17, 19, 21 and 23. Each cluster is generated six times with different value seed 1, 2, 5, 10, 20 and 100 (Michael et al, 2005). Cluster with the seed that given the smallest error is selected. Table I summarizes the results for all clusters.

TABLE I. Table Type Styles

Cluster Size	Seed	No of iterations	Squared errors	Has all attribute	Event Distribution
5	1	3	1200.0	No	No
10	20	4	966.0	Yes	Yes
13	1	3	917.0	Yes	Yes
15	20	4	869.0	Yes	Yes
17	20	4	828.0	Yes	Yes

19	20	5	807.0	Yes	Yes
21	20	3	781.0	Yes	Yes
23	10	4	744.0	Yes	Yes

Based on Table I, cluster 23 achieved the lowest error compared to the other clusters and all attributes are covered. Therefore cluster 23 is chosen to describe the distribution of the data in the dataset. Cluster 23 is then speared according to the students majoring. The aim is to investigate the category and criteria of students that are placed at the organization. Five new clusters have been found as shown in Table II. Table II shows that students with criteria as in clusters0, 6, 16 and 22 are mapped to private organizations while cluster12 mapped with goverment_owned organization. Appendix summarizes the result from the cluster analysis with cluster size of 23 based on organization classes.

TABLE II. CLUSTERS FOR ORGANIZATIONS

Cluster No	Major	City1	CGPA	Gender	Race	Organization	Clustered Instances
Cluster0	Software Engineering	N_Region1	CLASS4	F	India	Private	12 (2%)
Cluster6	Software Engineering	N_Region1	CLASS2	F	Malay	Private	42 (7%)
Cluster12	Software Engineering	E_Region3	CLASS3	M	Malay	Government_ owned	8 (1%)
Cluster16	Software Engineering	W_Region2	CLASS2	M	Malay	Private	17 (3%)
Cluster22	Software Engineering	N_Region1	CLASS1	F	Chinese	Private	14 (2%)

CONCLUSION

In this study data mining techniques namely k-means were used to mined hidden information from the large set of data. The generated clusters reveal pattern of decision that was performed by the practicum coordinator during the previous practicum placement process. This information can be used to support current student placement that is to match between student and the appropriate organization.

REFERENCES

[1] S. Hassan, W.H. Wan Ishak, M. A. Yahya and A.R. Chik, "Sistem Penempatan Pelajar Praktikum Atas Talian Sebagai Gerbang Jaringan Universiti-Industri", National Seminar on e-Community, 2009

[2] A.M. Ahmed, N.M. Norwawi and W.H. Wan Ishak, "Identifying Student and Organization Matching Pattern Using Apriori Algorithm for Practicum Placement", Proceedings of International Conference on Electrical Engineering and Informatics, 5-7 August 2009, Selangor, Malaysia, 2009, pp. 28-31

[3] A. Goswami, R. Jin, and G. Agrawal, "Fast and Exact Out-of-Core K-Means Clustering", Proceedings of the Fourth IEEE International Conference on Data Mining, 2004.

[4] J.Z.C. Lai, T.J. Huang, and Y.C. Liaw, "A fast-means clustering algorithm using cluster center displacement", Pattern Recognition, vol. 42(11), 2009, pp. 2551-2556.

[5] F. Mart, A. Troncoso, J.C. Riquelme, and J. M. Riquelme, "Partitioning-clustering techniques applied to the electricity price time series", Proceedings of the 8th international conference on Intelligent data engineering and automated learning. Birmingham, UK, 2007.

[6] J. Han, M. Kamber, and J. Pei, Data Mining, Second Edition: Concepts and Techniques. Ed. Elsevier Science, 2006

[7] L. Taoying, and C. Yan, "An improved k-means algorithm for clustering using entropy weighting measures", 7th World Congress on Intelligent Control and Automation, 2008, pp.149-153.

APPENDIX

Organization	Region	Criteria (K-means)

Government	N_Region1	Major= INFORMATION MANAGEMENT Or
		Major= NETWORKING
		CGPA= 2.5 -2.99
		Gender=Female
		Race= Malay

	W_Region2	Major= MULTIMEDIA
		CGPA= 2.5 -2.99
		Gender=Female
		Race= Malay Or Race= Chinese

		Major= INFORMATION MANAGEMENT
		CGPA= 2.5 -2.99
		Gender=Female
		Race= Malay

	S_Region4	Major= ARTIIFICIAL INTELLIGENCE
		CGPA= 3.0 – 3.49
		Gender=Female

		Race= Chinese
		
		Major= MULTIMEDIA
		CGPA= 2.5 -2.99
		Gender=Male
		Race= Malay

Government owned	N_Region1	Major= MULTIMEDIA
		CGPA= 2.5 -2.99
		Gender=Female
		Race= Malay
		
		Major= NETWORKING
		CGPA= 3.0 – 3.49
		Gender=Male
		Race= Chinese

	W_Region2	Major= ARTIIFICIAL INTELLIGENCE
		CGPA= 3.0 – 3.49
		Gender=Female
		Race= Malay

	E_Region3	Major= SAFTWARE ENGINEERING
		CGPA= 2.5 -2.99
		Gender=Male
		Race= Malay

	S_Region4	Major= NETWORKING
		CGPA= 2.5 -2.99
		Gender=Male
		Race= Chinese

Private	N_Region1	Major= SAFTWARE ENGINEERING
		CGPA= 2.5 – 2.99 , 3.0 – 3.49, 3.5 – 4.0
		Gender=Female
		Race= Malay or Indian ,Chinese
		
		Major= MULTIMEDIA
		CGPA= 2.5 – 2.99
		Gender=Female
		Race= Chinese
		
		Major= NETWORKING
		CGPA= 2.5 – 2.99
		Gender=Male
		Race= Chinese

	W_Region2	Major= SAFTWARE ENGINEERING
		CGPA= 3.0 – 3.49
		Gender=Male
		Race= Malay
		
		Major= MULTIMEDIA
		CGPA= 3.0 – 3.49
		Gender=Female
		Race= Malay Or Chinese

	E_Region3	Major= ARTIIFICIAL INTELLIGENCE
		CGPA= 2.5 – 2.99
		Gender=Male
		Race= Malay
		
		Major= MULTIMEDIA
		CGPA= 3.0 – 3.49
		Gender=Male
		Race= Malay
		
		Major= NETWORKING
		CGPA= 2.5 – 2.99
		Gender=Female
		Race= Malay