2009 International Conference on Electrical Engineering and Informatics
5-7 August 2009, Selangor, Malaysia

Identifying Student and Organization Matching Pattern Using Apriori Algorithm for Practicum Placement

Almahdi Mohammed Ahmed [1], **Norita Md Norwawi** [2], **Wan Hussain Wan Ishak** [3]

College of Arts & Sciences, *Universiti Utara Malaysia*
06010 Sintok, Kedah
[1] mhdi177@yahoo.co.uk
[2] nmn@uum.edu.my
[3] hussain@uum.edu.my

Abstract— Student's placement in industry for the practicum 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 association rules of data mining technique where Apriori algorithm was chosen. The data use consists of Bachelor of Information Technology and Bachelor in Multimedia students of University Utara Malaysia from the year 2004 till 2005. Two experiments were conducted using undirected data and directed data. The pattern extracted gave information on the previous matching process done by University –Industry Linkage Centre of University Utara Malaysia.

Keywords— **Data mining, KDD, association rules, apriori algorithm**

1.0 INTRODUCTION

The University-Industry Link (UIL) or previously known as Practicum Centre (PPR) of University Utara Malaysia (UUM) is responsible with the placement of students in the industry for the internship program. It is experiencing difficulty in matching organization‘s requirement with students profile for several reasons. This situation could lead to a mismatched between organization's requirement and students' background. Hence, students will face problems in giving good service to the company. On the other hand, companies too could be facing difficulties in training the students and assigning them with a project.

The placement must be based on certain criteria in order to best serve the organization and student. For example, 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. It has been decided by the UIL's top management that practicum students' should match the organization's requirement. However, due to the large number of students registered every semester, matching the organization with the students is a very tedious process.

The current procedures in matching organization and students 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's majoring. Usually, organization will request student with a specific majoring details. Other criterion is student's CGPA. Also, due to certain work prospect, some organization request student based on the gender and race. These criteria have been considered by the program coordinator in the placement process to ensure the right student being sent to the right organization.

This study aim to identify the patterns in matching organization and student and to extract hidden information from previously matched practicum placement datasets. This paper discusses the application of data mining technique particularly association rules to extract the historical placement pattern. This pattern will be a useful guideline for future organization and student matching. Data for the project is obtained from UUM's UIL. The data consist of all information technology (IT) and multimedia undergraduate students from the year 2004 till 2005. WEKA software is used to generate association rules.

2.0 LITERATURE REVIEW

Data mining have been applied in various research works. One of the popular techniques used for mining data in KDD for pattern discovery is the association rule [1]. According to [2] an association rule implies certain association relationships among a set of objects. It attracted a lot of attention in current data mining research due to it's capability of discovering useful patterns for decision support, selective marketing, financial forecast, medical diagnosis and many other application. The association rules technique works by finding all rules in a database that satisfies the determined minimum support and minimum confidence [3].

978-1-4244-4913-2/09/$25.00 ©2009 IEEE

An algorithm for association rule induction is the Apriori algorithm, proven to be one of the popular data mining techniques used to extract association rules [4], implemented the Apriori algorithm to mine single-dimensional Boolean association rules from transactional databases. The rules produced by Apriori algorithm makes it easier for the user to understand and further apply the result.

[5] Employed the association rule method specifically Apriori algorithm for automatically identifying new, unexpected, and potentially interesting patterns in hospital infection control. Another study by [6] employed Apriori algorithm to generate the frequent item sets and designed the model for economic forecasting. [7] Presented their methods on modelling and inferring user's intention via data. Experimental results showed that their proposed approach achieved 84% average accuracy in predicting user's intention, which is close to the precision (92%) of human prediction.

3.0 PATTERN EXTRACTION

In this study the KDD process which involved several steps: Selection, Pre-processing, Transformation, Pattern Extraction using data mining, and Interpretation/Evaluation [8].

3.1 Selection

Data used in this study was obtained from UIL database. These data have been generated by different reports among others Registered Students Report, Students' CGPA Report, Students' List Based on City Report. This data include all 2004 and 2005 Bachelor in Information Technology (BIT) and Bachelor in Multimedia (BMM) students. The initial data contains the performance profile gathered from a number of 998 students with 20 listed attributes which include Metric Number, Programme, Session, Major, Program Code, City1, City2, Address, 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 are *Majoring, CGPA, Gender, City1, Race, Organization* and *City3* chosen based on the suitability of the condition of the problems being discussed. The data were then processed for generating rules.

3.2 Pre-processing

Upon initial examination on the data, missing values of the attributes *City1, CGPA, Race, Gender, Organization* and *City3* were found and removed according to the numbers of missing values in one instance as part of the data cleansing process.

3.3 Transformation

According to [9], after the cleansing process, data is converted into a common format to make sure 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 KELAS4
2. If the *CGPA* = 2.5 Till 2.99 THEN Replace *CGPA* by KELAS3
3. If the *CGPA* = 3.0 Till 3.49 THEN Replace *CGPA* by KELAS2
4. If the *CGPA* = 3.5 Till 4.00 THEN Replace *CGPA* by KELAS1

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). 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 was than ready to be mined using association rules.

3.4 Pattern Extraction using Apriori Algorithm

In this study, the association rules using Apriori Algorithm was applied to the data for generating rules using WEKA software.

3.5 Interpretation/ Evaluation

During the process of pattern extraction, the acceptance of the output produced was evaluated in terms of accuracy and converge. This is to make sure that the generated rules are reliable and accurate. The accuracy of rules was obtained according to the value of confidence parameter determined earlier in the study while the degree of rules coverage was shown through the value of support parameter.

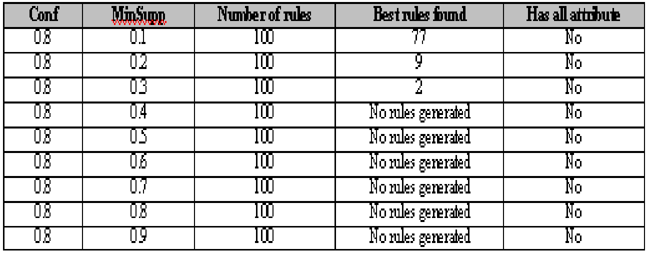
4.0 EXPERIMENTS AND RESULTS

An experiment was conducted to analyze the hidden relationship in the data using Apriori algorithm based on the organization category. Association analysis involves two experiments based on undirected data and directed data.

4.1 Results from Uundirected data

The first experiment was conducted based on the undirected data using Apriori algorithm. The undirected data is one that has been pre-processed but have not being grouped. The number of rules to be extracted is set to 100, 300 and 500 rules. In addition the value of confidences set to 0.8 as suggested by [7]. The minimum support used is from 0.1 until 0.9. The result from the first experiment (with 100 rules) is shown in Table 1.

Table 1: Experiment with 100 rules



Conf	MinSupp	Number of rules	Best rules found	Has all attribute
0.8	0.1	100	77	No
0.8	0.2	100	9	No
0.8	0.3	100	2	No
0.8	0.4	100	No rules generated	No
0.8	0.5	100	No rules generated	No
0.8	0.6	100	No rules generated	No
0.8	0.7	100	No rules generated	No
0.8	0.8	100	No rules generated	No
0.8	0.9	100	No rules generated	No

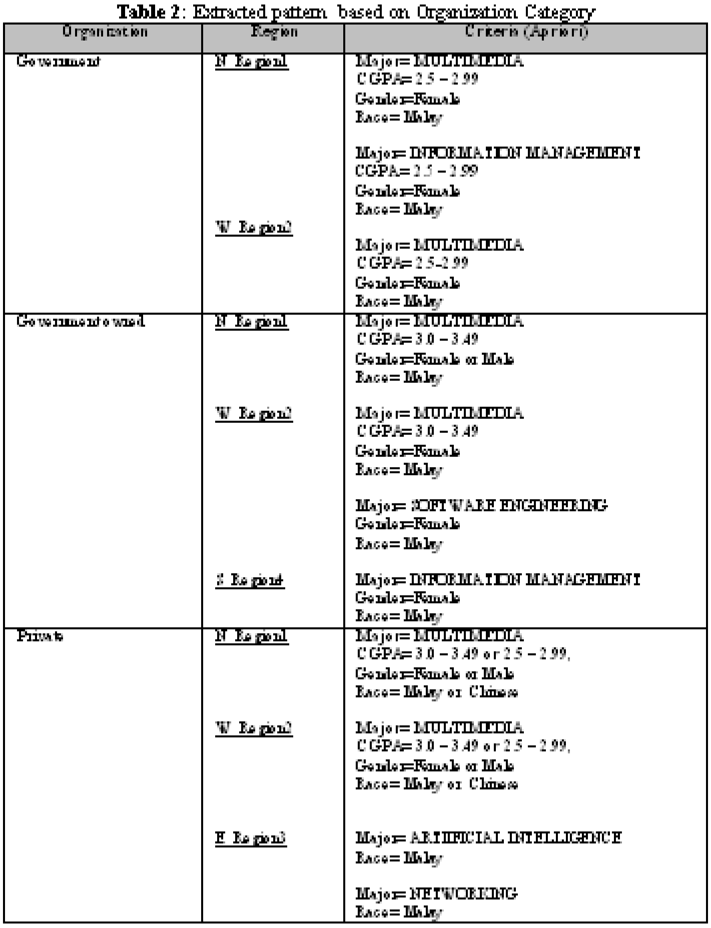
Table 1 shows that confidence value (*Conf*) is 0.8 with min support (*MinSupp*) 0.1 produce only 77 rules, which are far from the target (100 rules). Table 1 also shows that increasing the *minSupp* value does not improve the number of rules generated. The second experiment was then conducted by changing the number of rules to 300 and later to 500. The results showed no improvement similar to previous experiment.

From this experiment it was difficult to see any interesting pattern from rules generated which did not cover all attributes and instances. These rules might lead to wrong conclusion and decision.

4.2 Results from Directed data

In this experiment, the data has been grouped into three groups based on the Organization category. Again, the experiment was conducted using Apriori algorithm with the same specifications. Table 2 shows the results generated by Apriori for all three categories of organizations.

Table 2: Extracted pattern based on Organization Category



Organization	Region	Criteria (Apriori)

Government	N Region1	Major= MULTIMEDIA
		CGPA= 2.5 – 2.99
		Gender=Female
		Race= Malay
		
		Major= INFORMATION MANAGEMENT
		CGPA= 2.5 – 2.99
		Gender=Female
		Race= Malay

	W Region2	Major= MULTIMEDIA
		CGPA= 2.5-2.99
		Gender=Female
		Race= Malay

Government owned	N Region1	Major= MULTIMEDIA
		CGPA= 3.0 – 3.49
		Gender=Female or Male
		Race= Malay

	W Region2	Major= MULTIMEDIA
		CGPA= 3.0 – 3.49
		Gender=Female
		Race= Malay
		
		Major= SOFTWARE ENGINEERING
		Gender=Female
		Race= Malay

	S Region4	Major= INFORMATION MANAGEMENT
		Gender=Female
		Race= Malay

Private	N Region1	Major= MULTIMEDIA
		CGPA= 3.0 – 3.49 or 2.5 – 2.99,
		Gender=Female or Male
		Race= Malay or Chinese

	W Region2	Major= MULTIMEDIA
		CGPA= 3.0 – 3.49 or 2.5 – 2.99,
		Gender=Female or Male
		Race= Malay or Chinese

	E Region3	Major= ARTIFICIAL INTELLIGENCE
		Race= Malay
		
		Major= NETWORKING
		Race= Malay

4.3 Discussion on the Apriori result

From the pattern extracted, it was found that Apriori algorithm could generate patterns that are believed to be the factors that affect the matching process. From the experiment, extraction of the hidden information reveals that organization requirement can be fulfilled based on only three or four criteria. The best rules were selected where the Organization was set as the target of the students. The rules were evaluated based on the confidence and support. The best rules were chosen when the confidence is 90.0 % and the support also shown good support of 90.0 % and above.

Upon examining Table 2, example of pattern extracted are

IF students are from the Multimedia Majoring **AND**
Their CGPA is between 3.0 – 3.49 **AND**
They are Malay

THEN

The students were placed in the Northern Region and

In an Government Owned Organization

IF students are from the Artificial Intelligence and Networking Majoring **AND**

They are Malay

THEN

The students were placed in the Eastern Region and

In a Private Organization

5.0 CONCLUSION

This study has been implemented and conducted on existing data from UIL. In this study data mining techniques namely association rule was used to achieve the goal and extract the patterns from the large set of data. Using organization category as the target, the patterns extracted can provide information of the practicum placement and how the matching of the organization's requirement and student's criteria was done previously. Further analysis can be done by changing the target attributes.

REFERENCES

1. Hipp, J., Guntzer, U., Gholamreza, N. (2000). Algorithm for Association Rule Mining: A General Survey and Comparison, ACM SIGKDD, volume 2 (Issue 1), p. 58.
2. Fayyad, U. M., Shapiro, G. P., Smyth, P., and Uthurusamy, R. (1996). Advances in Knowledge Discovery and Data Mining, Cambridge, AAAI/MIT press
3. Liu, B., Hsu, W., Ma, Y. (1998). Integrating Classification and Association Rule Mining, American Association for Artificial Intelligence
4. Agrawal, R., C. Faloutsos, and A. N. Swami (1994). Efficient similarity search in sequence databases. In D. Lomet (Ed.), *Proceedings of the 4th International Conference of Foundations of Data Organization and Algorithms (FODO)*, Chicago, Illinois, pp. 69-84. Springer Verlag
5. Ma, Y., Liu, B., Wong, C. K., Yu, .S., & Lee, S. M. (2000). Targeting the Right Student Using Data Mining , ACM, PP. 457-463.
6. Sarjon Défit, Mohd Noor Md Sap. "An Economic Forecasting Based on Association Roles and Neural Network." Jurnal Teknologi Maklumat Jilid 13, no. Bil. 1 (Jun 2001): 42-55.
7. Chen, M. S., Han, J., and Yu, P. S. (1996). Data Mining: An Overview from a Database Perspective, IEEE Transaction on Knowledge and Data Engineering, pp. 866-883.
8. Jiawei Han, Micheline Kamber. "Data Mining : Concepts and Techniques " book: Data mining (2001).
9. Zhigang Li, Margaret H. Dunham, Yongqiao Xiao: STIFF: A Forecasting Framework for SpatioTemporal Data. Revised Papers from MDM/KDD and PAKDD/KDMCD 2002: 183-198