ISSN:0975-9646
D.Magdalene Delighta Angeline et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (6) , 2011, 2788-2791

Efficient Apriori Mend Algorithm for Pattern Extraction Process

D.Magdalene Delighta Angeline[1] ,I.Samuel Peter James[2]

[1], *Department of Computer Science and Engineering,*
Dr.G.U.Pope College of Engineering, Sawyerpuram-62825, Tamilnadu, India.
[2], *Department of Computer Science and Engineering,*
Chandy College of Engineering,Thoothukudi, Tamilnadu, India.

Abstract— **Association rules reflect the inner relationship of data. Discovering these associations is beneficial to the correct and appropriate decision made by decision-makers. The association rules provide an effective means to found the potential link between the data, reflecting a built-in association between the data. In this paper, we make an in-depth the study of the mining association rules for Student's placement in industry. Student's placement 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 using Apriori Mend algorithm.**

Keywords— **Apriori Mend Algorithm, Association rules, Data mining, Knowledge Discovery**

I. INTRODUCTION

Usually, the problem of Mining Association rules can be divided into two phases: Phase one, developed by the users in accordance with the minimum degree of support from the database to find a frequency greater than or equal to the minimum support of all frequent item sets; phase two, it is generated by the first phase of the project set to produce frequent association rules. The earliest in 1993 Agrawal, proposed AIS algorithm, the AIS algorithm has too many sets of candidate projects, which result in the last Mining Association Rules poor efficiency of the way, so in 1994, Agrawal, also filed a Apriori algorithm. Later many Apriori based on the improved algorithm have been proposed, such as Savasere, who proposed Partition algorithm, Toivonen made Sampling algorithm, Park, who proposed the use of technology DHP hash algorithm in 1995, on the above-mentioned study, the mining association rules is to increase the efficiency of its methods not only reducing the non-collection of related items.

This paper is to study data mining in the mining of association rules to extract the historical placement pattern with Apriori Mend algorithm which will be helpful for the industry in the placement. The teaching organization 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 Chennai should not be sent to an organization located in Bangalore. This is to avoid problems in terms of accommodation, financial, and social. It has been decided 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 organizations 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 Percentage. 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 also aims to identify the patterns in matching organization and student and to extract hidden information from previously matched practicum placement datasets. The problem of finding association rules falls within the purview of database mining, also called knowledge discovery in databases.

II. 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 its capability of discovering useful patterns for decision support, selective marketing, financial forecast, medical diagnosis and many other applications. The association rules technique works by finding all rules in a database that satisfies the determined minimum support and minimum confidence [3].

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 employed Apriori algorithm to generate the frequent item sets and designed the model for economic forecasting, presented their methods on modelling and inferring user's intention via data. The Apriori algorithm is improved for their efficiency and performance of the database.

III. PATTERN EXTRACTION

By analyzing a large amount of data a pattern or rule is extracted.

A. Selection

The data have been generated by different reports among others Registered Students Report, Students' Mark Report, Students' List Based on City Report. The initial data contains the performance profile gathered from a number of 125 students with 20 listed attributes which include Register Number, Programme, Duration, Program Code, City1, City2, Address, Address State, Percentage, Gender, Race Code, Race, Organization, Address1, Address2, 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 two categories (Government and Private). Based on the discussion with the program coordinator, all 125 data are used in this study. The selected attributes are Majoring, Percentage, 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.

B. Pre-processing

Now upon initial examination on the data, missing values of the attributes City1, Percentage, 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.

C. 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 *Percentage* to string data.

1. If the *Percentage* = 81 Till 90 THEN Replace *Percentage* by S1
2. If the *Percentage* = 75 Till 80 THEN Replace *Percentage* by S2
3. If the *Percentage* = 70 Till 74 THEN Replace *Percentage* by S3
4. If the *Percentage* = 65 Till 69 THEN Replace *Percentage* by S4

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, KODAMPAKAM and GUINDY have the same code 02 then they were converted into one Region (N_Region). Organization's name was also transformed by into two categories (Government and Private). After all pre-processing and transformation have been implemented, the data was than ready to be mined using association rules.

D. Pattern Extraction using Apriori Mend Algorithm

In this study, the association rules using Apriori Mend algorithm was applied to the data for generating rules.

E. Apriori Mend Algorithm

Apriori Mend algorithm upgrade the efficiency of the project set and then use these to identify the frequent item sets derived association rules. Apriori Mend algorithm mainly composed of five steps:

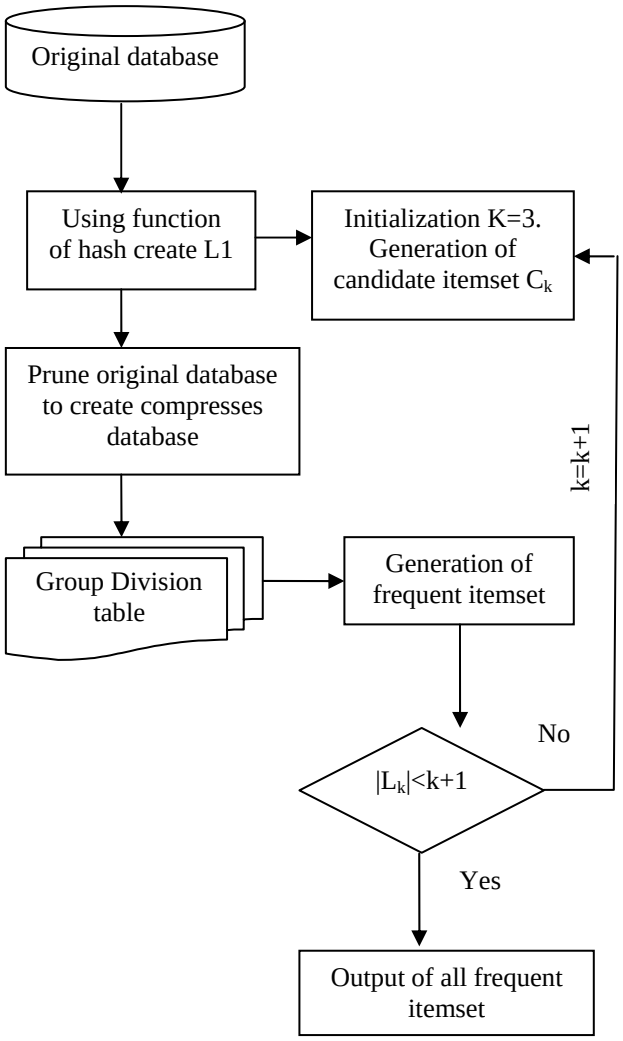

Fig. 1 Flowchart for Apriori Mend Algorithm

1. A frequent 2-searchable database use perfect hash function and the user to specify the minimum support to generate frequent-2 itemset, the steps must pay attention to the smallest degree of support for the choice, because if the specified minimum support, which will cut off too much data; otherwise might have left too much data to be a significant amount of storage space.
2. Frequent use of cutting-2 in the transactions of the database, many of the transactions or the entire transaction records is of no use, so in scanning the database, one step at the same time using the frequent two-item database of pruning, Transactions will be recorded in the database does not belong to L2 (item) to delete the item.
3. According to records of transactions and length of division in order to form into groups to quickly, find all the frequent item sets, and further

deduced all the association rules, the frequent use of set of pruning database, the data will be pruned based on records of transactions length division and deposited in the corresponding division in the table, its approach to K projects will include records of transactions to the division for the K-K and deposited in a division table.

4. A candidate implementation procedures must be set by the frequent-2 item set, the expansion of layers, take advantage of the K-1 layer received frequent item sets combined, introduced the use of front operations and do not need pruning which can have a direct candidate K. The algorithm is used to produce the K-item collection of candidates CK.
5. A frequent item sets in the calculation of the length of the candidate itemset for the K-degree of support, because of its length K, it just needs from the first division table K-start, the removal of each group in the table service records, according to records in the affairs CK filtering have CK", calculated each services include CK" set of items, a subgroup in the search end immediately after calculating its support, and do not necessarily have to wait until all the data until after the scan, which can immediately determine whether the degree of support by setting the minimum support, if you can meet immediately become a frequent item sets, this candidate will not collect items need to continue with the comparison to the contrary to section K +1 form a group to compare, until its support with minimum support, or has a subgroup compared to the last table so far. When LK generated after the calculation | LK |, if | LK | <k +1, the end of the algorithm. Otherwise, return to step 4.

G 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.

IV EXPERIMENTS

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 1 shows the results generated by Apriori Mend algorithm for all two categories of organizations.

A 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.

TABLE 1
EXTRACTED PATTERN BASED ON ORGANIZATION CATEGORY

Organization	Region	Criteria (Apriori)
Government	N_Region1	Major=Computer Science and Engineering Percentage=75-80 Gender=Male Race = Guindy
		Major=Electronics Communication and Engineering Percentage=75-80 Gender=Male Race = Guindy
	W_Region2	Major=Electrical and Electronics Engineering Percentage=75-80 Gender=Male Race = Guindy
Private	N_Region1	Major=Computer Science and Engineering Percentage=70-74 or 75-80 Gender=Female or Male Race = Guindy or Kodambakam
	W_Region2	Major=Electronics Communication and Engineering Percentage=70-74 Gender=Male Race = Guindy
		Major=Electrical and Electronics Engineering Percentage=70-74 Gender=Male Race = Guindy

Fig. 2 Generation of association rules

Figure 2 shows the generation of association rules with minimum support count. Upon examining Table 1, example of pattern extracted is:

IF students are from the Computer Science and Engineering
AND
Their Percentage is between 75-80 AND
They are Guindy
THEN
The students were placed in the Northern Region and
In an Government Organization

IF students are from the Electrical and Electronics Engineering and Electronics Communication and Engineering
Majoring AND
Their Percentage is between 70-74 AND
They are Guindy
THEN
The students were placed in the Western Region and
In a Private Organization

V CONCLUSIONS

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. Apriori Mend algorithm based on the structure of Apriori algorithm, uses the default minimum support to prune the database itemset, deleting the needless itemset. And then grouped the pruned database according to the transaction length, establishing a sub-groups tables to meet the group table quickly find all the characteristics of the frequent item sets. Finally, Apriori Mend algorithm is found to be more excellent than the traditional method Apriori algorithm in the efficiency of performance.

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] R. Agrawal, T. Imielinski, and A.Swami.Database mining: A performance perspective.IEEE Transactions on Knowledge and Data Engineering, 5(6):914{925, December 1993. Special Issue on Learning and Discovery in Knowledge-Based Databases.

[7] Almahdi Mohammed Ahmed , Norita Md Norwawi , Wan Hussain Wan Ishak(2009), Identifying Student and Organization Matching Pattern Using Apriori Algorithm for Practicum Placement, *International Conference on Electrical Engineering and Informatics ,Selangor, Malaysia*.

[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.

[10] Rakesh Agrawal Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules, Proceedings of the 20th VLDB Conference Santiago, Chile (1994).

D.Magdalene Delighta Angeline is a Lecturer in the Department of Computer Science and Engineering in Dr.G.U.Pope College of Engineering, Sawyerpuram, Tamilnadu, India. She obtained her Bachelor degree in Information Technology from Anna University, Chennai in the year 2007 and she obtained her Master degree in Computer and Information Technology in Manonmaniam Sundaranar University, Tirunelveli. She has over 4 years of Teaching Experience and published six paper in national conferences and one in International conferences. Her current area of research includes Image Processing, Neural Networks, and Data Mining. Email: magdalenedelighta@gmail.com

I. Samuel Peter James is a Lecturer in the Department of Computer Science and Engineering in Chandy College of Engineering, Thoothukudi, Tamilnadu, India. He obtained his Bachelor degree in Computer Science and Engineering from Anna University, Chennai in the year 2009 and he is doing his Master degree in Computer Science and Engineering and also doing M.B.A. in Manonmaniam Sundaranar University, Tirunelveli. He has over 2.2 years of Teaching Experience His current area of research includes Image Processing, Neural Networks, and Data Mining. Email: i.samuelpeterjames@gmail.com