Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 28 (2011) 107 – 111

Procedia
Social and Behavioral Sciences

www.elsevier.com/locate/procedia

WCETR 2011

Extracting highly positive association rules from students' enrollment data

Zailani Abdullah[a,*], Tutut Herawan[b], Noraziah Ahmad[b], Mustafa Mat Deris[c]

[a]*Department of Computer Science, Universiti Malaysia Terengganu 21030 Kuala Terengganu, Terengganu, Malaysia*
[b]*Faculty of Computer System and Software Engineering, Universiti Malaysia Pahang Lebuhraya Tun Razak, 26300 Kuantan Pahang, Malaysia*
[c]*Faculty of Computer Science and Information Technology Parit Raja, Batu Pahat 86400, Johor, Malaysia*

Abstract

Association Rules Mining is one of the popular techniques used in data mining. Positive association rules are very useful in correlation analysis and decision making processes. In educational context, determine a "right" program to the students is very unclear especially when their chosen programs are not selected. In this case, normally they will be offered to other programs based on the programs' availability and not according to their program's field interests. The main concern is, by assigning inappropriate program which is not reflected their overall interest; it may create serious problems such as poorly in academic commitment and academic achievement. Therefore, Therefore in this paper, we proposed a model which consists of pre-processing, mining patterns and assigning weight to discover highly positive association rules. We examined the previous chosen programs by computer science students in our university for July 2008/2009 intake. The result shows that the proposed model can mine association rules with high correlation. Moreover, for data analysis, there are existed students that have been offered in computer science program at our university but not within their program's field interests.

Keywords: Positive association rules; Educational; Progams.

1. Introduction

Data mining is referred to the process of extracting hidden and useful information in large data repositories (Tan *et al*., 2005). It is an integral part of knowledge discovery in databases (KDD) and currently seen as an important media for details information analysis in many research applications. One of the emerging interdisciplinary research areas (Baker and Yacef, 2009) in educational context is educational data mining. Educational data mining can be defined as the application of data mining techniques to educational data (Romero *et. al*, 2008). The main concern of educational data mining is to develop methods that can explore the interesting information extracted from educational settings, and employ those methods to better understand the students, and the settings in which they learn (http://www.educationaldatamining.org). One of the most important issues in data mining application is

* Zailani Abdullah. Tel.: +609-6683536; fax: +609-6694660.
E-mail address: zailania@umt.edu.mv

1877-0428
doi:10.1016/j.sbspro.2011.11.022

association rules mining. The problem of association rules mining was first encountered by Agrawal *et al.* (1993) for the purpose of market-basket analysis. In association rules, a set of item is also defined as an itemset. Least itemset is a set of item that is rarely found in the database. It is also known as non-frequent, unusual, exceptional, abnormal, in-balance or sporadic itemset. Discovering of such itemset is very important because it may reveals a valuable knowledge for certain domain applications such as detection for air pollution, network intruders, machine critical faulty (Abdullah *et al.*, 2010), abnormal learning problems (Romero *et al.*, 2010), and many more. From the previous works, most of the tradition association rules mining algorithms (Kiran *et al.*, 2009; Zhou and Yau, 2007; Koh *et al.*, 2005; Yun *et al.*, 2003; Liu *et al.*, 1999; Wang *et al.*, 2008; Tao *et al.*, 2003; Ding, 2005) are still have a limitation in term of efficiency, scalability and are rarely applied into the real datasets. The low minimum support can be set to capture the least itemset. However, the trade off is it may generate the huge number of association rules. As a result, it is enormously difficult to identify which association rules are most interesting and really significant. Furthermore, the low minimum support will also proportionally increase the computational performance and its complexity. Since the complexity of study, difficulties in algorithms (Yun *et al.*, 2003) and it may require excessive computational cost, there are very limited attentions have been paid to discover least association rules. Extracting a complete set of positive association rules (Yang *et al.*, 2009, Wu *et al.*, 2004) is very important especially in educational context. For example, in every July semester, our university receives approximately 160 students to enroll in computer science program. However, in the first meeting, many of them are not keen to study in this program. In Malaysia, all potential students are required to apply via online system to choose 8 preferred bachelor programs offered by Malaysian universities. Many students are not clear and always mixed up with the various fields of study. For example, the students might put together the contradict program's field such as social science, banking, forestry and computer. The question is how to discover program's field interests of the students since there is no information provided in online application. At the moment, if their chosen programs are not selected, they will be offered to any programs in the university without taking into consideration their program's field interests. The issue is, by offering inappropriate program which is not reflected their overall interest; it may create serious problems such as poorly in academic commitment and academic achievement. Therefore, this study is very important since it can reveal the degree of correlation between the pervious chosen programs. It can be used for the university representative as a guideline in offering the appropriate programs to them. Moreover, it can assist the university's policy maker to comprehend the issues and also enhance the current educational standards and managements process as a whole (Sevindik *et al.*, 2010; Buldu *et al.*, 2010; Romero *et al.*, 2009; Romero *et al.*, 2008; Enclieva *et al.*, 2006). In this paper, we propose a model which consists of data preprocessing, pattern tree construction and finally mining positive association rules. Here, lift measurement (Brin *et al.*, 1997) is employed in classifying the association rules. The experiment was performed based on the 2008/2009 intake students' offered in Bachelor of Information Technology (Software Engineering) at Universiti Malaysia Terengganu. Due to the confidentiality, the details about the dataset will not be disclosed.

The rest of the paper is organized as follows. Section 2 explains the basic concepts and terminology of association rule mining. Section 3 discusses the proposed method. This is followed by performance analysis in section 4. Finally, conclusion and future direction are reported in section 5.

2. Preliminaries

2.1. Association Rules

Throughout this section the set $I = \{i_1, i_2, \cdots, i_{|A|}\}$, for $|A| > 0$ refers to the set of literals called set of items and the set $D = \{t_1, t_2, \cdots, t_{|U|}\}$, for $|U| > 0$ refers to the data set of transactions, where each transaction $t \in D$ is a list of distinct items $t = \{i_1, i_2, \cdots, i_{|M|}\}$, $1 \le |M| \le |A|$ and each transaction can be identified by a distinct identifier TID. A set $X \subseteq I$ is called an itemset. An itemset with k-items is called a k-itemset. The support of an itemset $X \subseteq I$, denoted $\text{supp}(X)$ is defined as a number of transactions contain *X*. Let $X, Y \subseteq I$ be itemset. An association rule between sets *X* and *Y* is an implication of the form $X \Rightarrow Y$, where $X \bigcap Y = \phi$. The sets X and Y are called antecedent and consequent, respectively. The support for an association rule $X \Rightarrow Y$, denoted $\text{supp}(X \Rightarrow Y)$, is defined as a

number of transactions in D contain $X \bigcup Y$. The confidence for an association rule $X \Rightarrow Y$, denoted $\text{conf}(X \Rightarrow Y)$ is defined as a ratio of the numbers of transactions in *D* contain $X \bigcup Y$ to the number of transactions in D contain X. Thus $\text{conf}(X \Rightarrow Y) = \text{supp}(X \Rightarrow Y) / \text{supp}(X)$.

2.2. Correlation Analysis

A few years after the introduction of ARs, Aggrawal *et al.* (1993) and Brin *et al.* (1997) realize the limitation of the confidence-support framework. Many studies have shown that the confidence-support framework alone is insufficient at discovering the interesting ARs. Therefore, the correlation can be used as complimentary measure of this framework. This leads to correlation rules as

$$A \Rightarrow B \quad (\text{supp}, \text{conf}, \text{corr}) \tag{1}$$

The correlation rule is measure based on the minimum support, minimum confidence and correlation between itemsets *A* and *B*. There are many correlation measures applicable for ARs. One of the simplest correlation measures is Lift. The occurrence of itemset *A* is independence of the occurrence of itemset *B* if $P(A \cup B) = P(A)P(B)$; otherwise itemset *A* and *B* are dependence and correlated. The lift between occurrence of itemset *A* and *B* can be defined as:

$$\text{lift}(A, B) = \frac{P(A \cap B)}{P(A)P(B)} \tag{2}$$

The equation of (4) can be derived to produce the following definition:

$$\text{lift}(A, B) = \frac{P(B \mid A)}{P(B)} \tag{3}$$

or

$$\text{lift}(A, B) = \frac{\text{conf}(A \Rightarrow B)}{\text{supp}(B)} \tag{4}$$

The strength of correlation is measure from the lift value. If $\text{lift}(A, B) = 1$ or $P(B \mid A) = P(B)$ $(\text{or } P(A \mid B) = P(B))$ then *B* and *A* are independent and there is no correlation between them. If $\text{lift}(A, B) > 1$ or $P(B \mid A) > P(B)$ $(\text{or } P(A \mid B) > P(B))$, then A and B are positively correlated, meaning the occurrence of one implies the occurrence of the other. If $\text{lift}(A, B) < 1$ or $P(B \mid A) < P(B)$ $(\text{or } P(A \mid B) < P(B))$, then A and B are negatively correlated, meaning the occurrence of one discourage the occurrence of the other. Since lift measure is not down-ward closed, it definitely will not suffer from the least item problem. Thus, least itemsets with low counts which per chance occur a few times (or only once) together can produce enormous lift values.

3. Methodology

3.1. Preprocessing

Determine field (domain) of the program. A flat file contains program's field and program's code is produced. They are split by a blank space.
Extract the list of chosen programs. At this part, a flat file called "ChosenProgramData" is created. Each record in this file is separated by a blank space.
Extract the list of chosen fields. A flat file which contain only program's field is produced and named as "FPDataset". A blank space is used to differentiate between the program's fields.

3.2. Mining Patterns (Itemsets)

Determine Interval Support for least Itemset. An itemset is said to be least if the support count satisfies in a range of threshold values called Interval Support (ISupp).

Construct Significant Least Pattern Tree. A Significant Least Pattern Tree (SLP-Tree) is built only with the items that satisfy the ISupp.
Generate Significant Least Pattern Growth (SLP-Growth). SLP-Growth is an algorithm that generates significant least itemsets from the SLP-Tree by exploring the tree based on a bottom-up strategy.

3.3. Assigning Weight (Measurement)

Apply Correlation. The weighted association rules (value) are derived from the formula (4). This correlation formula is also known by lift.
Discovery Highly Correlated Least Association Rules. From the list of weighted association rules, the algorithm will begin to scan all of them.

4. Experiment Test

In this section, we do experiment tests with Lift measurements. The weight of all association rules have been assigned accordingly. We evaluate the proposed model to 2008/2009 intake students in computer science program. The data was obtained from Division of Academic, Universiti Malaysia Terengganu in a text file and Microsoft excel format. There were 160 students involved and their identities were removed due to the confidentiality agreement. In the original set of data, it consists of 35 attributes and the detail information were explained in 10 tables which provided in Microsoft excel format.
The 8 chosen programs by the students are extracted according to the fix location in the original flat file. There were in total of 822 bachelors programs offered in Malaysian public universities for July 2008/2009 students' intake. From this figure, 342 bachelor programs were selected by our 160 students and it can be generalized into 47 unique general fields. The total of 4,177 association rules was successfully extracted. Table 1 depicts top 10 of association rules based on the 3% of minimum support. From these association rules, only 3 rules are reasonable and acceptable by referring to the interest relationship between programs. The rest are quite contradiction in term of interest relationship between the programs being selected by the students.

Table 1. Top 10 of positive association rules

No.	Association rules	Support of Antecedent	Support of Consequence	Support of Itemset	Confidence	Lift
1	28 –> 40	68.75	3.75	3.12	100	116.50
2	25 12 --> 9	11.88	4.38	4.38	100	99.86
3	12 --> 9	14.38	4.38	4.38	100	99.86
4	25 --> 9	90.63	4.38	4.38	100	99.86
5	25 28 12 --> 9	7.50	4.38	3.12	100	99.86
6	28 12 --> 9	10.00	4.38	3.12	100	99.86
7	25 28 --> 9	60.63	4.38	3.12	100	99.86
8	28 --> 9	68.75	4.38	3.12	100	99.86
9	12 --> 11	14.38	5.00	3.75	100	87.38
10	45 --> 11	38.75	5.00	3.12	100	87.38

5. Conclusion

One of the famous techniques in data mining is association rules mining. From the extracted rules, positive association rules are consider very useful for correlation analysis and decision making processes. Nowadays, determination of the appropriate program for prospect students is very difficult and usually ends up with programs availability. This issue is raised up when the previous students' chosen programs are not selected and they are offered to the program which is not within their program's field interests. Therefore in this paper, a model which consists of pre-processing, mining patterns and assigning weight to discover highly positive association rules is

proposed. The students' enrolment data of computer science program (intake 2008/2009) at University Malaysia Terengganu is examined. The result shows that the proposed model can discover the association rules with highly correlated. From the extracted association rules, there are existed students that have been offered in computer science program at our university but not within their program's field interests.

In a near future, we are going to evaluate the proposed model to others datasets. We also believed that the model is also suitable to others real domain applications.

Acknowledgement

This work is supported by Universiti Malaysia Pahang under Short Term Grant Vote Project RDU 090306.

References

Abdullah, Z., Herawan, T., Deris, M.M. (2010) 'Mining Significant Least Association Rules using Fast SLP-Growth Algorithm', Lecture Notes in Computer Science, 6059, Springer, Berlin Heidelberg New York, pp.324–336.

Agrawal, R., Imielinski, T., and Swami, A. (1993) 'Database Mining: A Performance Perspective', IEEE Transactions on Knowledge and Data Engineering 5 (6), pp.914–925.

Baker, R., and Yacef, K. (2009), 'The state of educational data mining in 2009: a review and future visions', Journal of Education Data Mining. Volume 1, Issue 1, pp. 3–17

Brin,S.Motwani,R.Ullman,J.and Tsur,S. (1997), 'Dynamic Itemset counting and implication rules for market basket data', SIGMOD-97, 1997, pp.255-264.

Buldu, A., and Ucgun, K. (2010), 'Data mining application on students' data', Procedia Social and Behavioral Sciences 2 (2010) pp. 5251-5259.

Ding, J. (2005) 'Efficient Association Rule Mining among Infrequent Items', Ph.D. Thesis, n University of Illinois at Chicago

Enclieva, S., and Tumin, S. (2006) 'Application of association rules for efficient learning work-flow'. In FIP International Federation for Information Processing, Volume 228, Intelligent Information Processing III, eds. Z. Shi, Shimohara K., Feng D., (Boston: Springer), pp. 499-504.

Kiran, R.U., and Reddy, P. K. (2009) 'An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules'. In Proceeding of IEEE Symposium on Computational Intelligence and Data Mining, pp.340–347.

Koh, Y.S., and Rountree, N. (2005) 'Finding Sporadic Rules using Apriori-Inverse'. In: Ho, T.B., Cheung, D., Liu, H. (Eds): Advances In Knowledge Discovery And Data Mining: 9th Pacific-Asia conference (PAKDD 2005), Hanoi, Vietnam. Lecture Notes in Computer Science, 3518, Springer, Berlin Heidelberg New York, pp. 97–106.

Liu, B., Hsu, W., and Ma, Y. (1999) 'Mining Association Rules with Multiple Minimum Supports', In Proceeding of ACM SIGKDD'07, pp.337–341.

Liu, C.C (2010), 'The relationship between employees' perception of emotional blackmail and their well-being', WCPCG 2010, Procedia - Social and Behavioral Sciences, Volume 5, 2010, pp. 299-303

Romero, C., Romero, J.R., Luna, J. M., Ventura S. (2010), 'Mining rare association rules from e-Learning data', In Proceeding of The Third International Conference of Education Data Mining, Pittsburgh, USA, pp. 171-180

Romero, C., Venturaa, S., and Garcíaa, E (2008) 'Data mining in course management systems: Moodle case study and tutorial', Journal of Computers & Education, Volume 51, Issue 1, pp.368–384.

Romero, C., Venturaa, S., Zafraa, A., and Brab., P. (2009), 'Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems', Journal of Computers & Education Volume 53, Issue 3, November 2009, pp. 828-840

Sevindik, T., Demirkeser, N., and Cömert, Z., (2010), 'Virtual education environments and web mining', Procedia Social and Behavioral Sciences 2 (2010) pp. 5120–5124

Szathmary, L., Napoli, A., Valtchev, P. (2007), 'Towards rare itmeset mining', In International Conference on Tools with Artificial Intelligence, Washington, USA, pp. 305-312

Tan, P-N., Steinbach, M., and Vipin, K. (2006). Introduction to Data Mining. Addison-Wesley.

Tanimoto, S.L. (2007), 'Improving the prospects for educational data mining', In Proceedings of the complete on-line proceedings of the workshop on data mining for user modelling, 11th International conference on user modeling (UM 2007), pp. 106-110

Tao, F., Murtagh, F., and Farid, M. (2003) 'Weighted Association Rule Mining using Weighted Support and Significant Framework', In Proceeding of ACM SIGKDD '03, pp.661–666.

Wang, H, Zhang, X., and Chen, G (2008), 'Mining a Complete Set of Both Positive and Negative Association Rules from Large Databases', In Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD08), LNAI 5012, Springer-Verlag Berlin Heidelberg, Antwerp, Belgium, pp.777-784.

Wu, X, Zhang, C., and Zhang, S. (2004), 'Efficient Mining of Both Positive and Negative Association Rules', Journal ACM Transactions on Information Systems (TOIS), Volume 22 Issue 3, July 2004, pp.381–405

Yang, J., Zhao, C. (2009), 'Study on the Data Mining Algorithm Based on Positive and Negative Association Rules', Journal of Computer and Information Science, Vol. (2), No. (2), May 2009, pp. 103-106.

Yun, H., Ha, D., Hwang, B., and Ryu, K.H. (2003) 'Mining Association Rules on Significant Rare Data using Relative Support', The Journal of Systems and Software 67 (3), pp.181–19.

Zhou L., and Yau, S. (2007) 'Assocation Rule and Quantative Association Rule Mining among Infrequent Items'. In Proceeding of ACM SIGKDD'07, Article No. 9.

http://www.educationaldatamining.org/ Retrieved on 14-03-2011

http://www.mqa.gov.my/ Retrieve on 14-03-2011