Discovering Interesting Association Rules from Student Admission Dataset

Zailani Abdullah[1] Tutut Herawan[2], Mustafa Mat Deris[3]

[1]Department of Computer Science, Universiti Malaysia Terengganu
[2]Faculty of Computer Science & Information Technology, University Malaya
[2]Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
[3]Faculty of Computer Science & Information Technology, University Tun Hussein Onn Malaysia

zailania@umt.edu.my, tutut@um.edu.my, mmustafa@uthm.edu.my

Abstract. Finding the interesting rules from data repository is quite challenging weather for public or private sectors practitioners. Therefore, the purpose of this study is to apply an enhanced association rules mining method, so called SLP-Growth (Significant Least Pattern Growth) proposed by [11,36] to mining the interesting association rules based on the student admission dataset. The dataset contains the records of preferred programs being selected by post-matriculation or post-STPM students of Malaysia via Electronic Management of Admission System (e-MAS) for the year 2008/2009. The results of this study will provide useful information for educators and higher university authority personnel in the university to understand the programs' patterns being selected by them.

Keywords: Association rule mining, significant least patterns, students.

1 Introduction

Generally, public universities are among the ultimate directions for almost post-matriculation or post-STPM students in Malaysia. After the students obtaining the actual result of the examination, they have to choose their preferred programs at Malaysian public universities via Electronic Management of Admission System (e-MAS). The issue is, for the average and the lower grades students, they might not be offered to their preferred programs. Based on this situation, many studies [1-3] have been carried out to ensure prolong of the students at university. Currently, there is an increasing interest in data mining and educational systems, making educational data mining as a new growing research community [4]. One of the popular data mining methods is Association Rules Mining (ARM) [5]. It aims at discovering the interesting correlations, frequent patterns, associations or casual structures among sets of items in the data repositories. The problem of association rules mining was first introduced by Agrawal for market-basket analysis [6,7,8]. After the introduction of Apriori [6], many studies [13-26] have been done pertinent to Association Rules (ARs). Generally, an item is said to be frequent if it appears more than a minimum support threshold. Least

T. Herawan et al. (eds.), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, Lecture Notes in Electrical Engineering 285, DOI: 10.1007/978-981-4585-18-7_16,

item is an itemset whose rarely found in the database but it may produce interesting and useful ARs. In this paper, we employ SLP-Growth algorithm and Critical Relative Support (CRS) measure [11,36] to capture interesting rules from student admission dataset. The dataset was taken from Division of Academic, Universiti Malaysia Terengganu for 2008/2009 intake students in computer science program. The results of this study will provide useful information for educators or higher university personnel authority to offer more relevant programs to the potential students rather than by chance or unguided technique.

The reminder of this paper is organized as follows. Section 2 describes the related work. Section 3 describes the essential rudiments. Section 4 describes the employed method, SLP-Growth algorithm. This is followed by performance analysis through student admission dataset in section 5. Finally, conclusion of this work is reported in section 6.

2 Related Works

For the past decades, there are several efforts has been made to discover the interesting ARs. Zhou *et al.* [27] suggested a method to mine the ARs by considering only infrequent itemset. Ding [28] proposed Transactional Co-occurrence Matrix (TCOM for mining association rule among rare items. Yun *et al.* [9] introduced the Relative Support Apriori Algorithm (RSAA) to generate rare itemsets. Koh *et al.* [29] suggested Apriori-Inverse algorithm to mine infrequent itemsets without generating any frequent rules. Liu *et al.* [30] proposed Multiple Support Apriori (MSApriori) algorithm to extract the rare ARs. From the proposed approaches [9,28–30], many of them are using the percentage-based approach to improve the performance as faced by the single minimum support based approaches. In term of measures, Brin *et al.* [31] introduced objective measure called lift and chi-square as correlation measure for ARs. Lift compares the frequency of pattern against a baseline frequency computed under statistical independence assumption. Omiecinski [32] proposed two interesting measures based on downward closure property called all confidence and bond. Lee *et al.* [33] suggested two algorithms for mining all confidence and bond correlation patterns by extending the pattern-growth methodology Han *et al.* [34]. In term of mining algorithms, Agrawal *et al.* [6,7] proposed the first ARs mining algorithm called Apriori. Han *et al.* [35] suggested FP-Growth algorithm which amazingly can break the two limitations as faced by Apriori series algorithms. Recently, Educational Data Mining (EDM) has emerged as an important research area in order to resolve educational research issues [37]. Kumar *et al.* [38] enhanced the quality of students' performances at post graduation level via association rule mining. Garcial *et al.* [39] described a collaborative educational data mining tool based on association rule mining for the ongoing improvement of e-learning courses. Tair *et al.* [40] used association rule mining technique to analyze to improve graduate students' performance, and overcome the problem of low grades of graduate students. Chandra *et al.* [41], applied the association rule mining technique to identifies the students' failure patterns in order to improve the low capacity students' performances.

3 Essential Rudiments

3.1 Association Rules (ARs)

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.

Definition 1. *A set $X \subseteq I$ is called an itemset. An itemset with k-items is called a k-itemset.*

Definition 2. *The support of an itemset $X \subseteq I$, denoted $\mathrm{supp}(X)$ is defined as a number of transactions contain X.*

Definition 3. *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.*

Definition 4. *The support for an association rule $X \Rightarrow Y$, denoted $\mathrm{supp}(X \Rightarrow Y)$, is defined as a number of transactions in D contain $X \bigcup Y$.*

Definition 5. *The confidence for an association rule $X \Rightarrow Y$, denoted $\mathrm{conf}(X \Rightarrow Y)$ 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*

$$conf(X \Rightarrow Y) = \frac{\sup\mathrm{p}(X \Rightarrow Y)}{\sup\mathrm{p}(X)}.$$

ARs that satisfy the minimum support and confidence thresholds are said to be strong.

4 Methodology

4.1 Algorithm Development

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).

The Interval Support is a form of ISupp (ISMin, ISMax) where ISMin is a minimum and ISMax is a maximum values respectively, such that $\text{ISMin} \geq \phi$, $\text{ISMax} > \phi$ and $\text{ISMin} \leq \text{ISMax}$.

Construct Significant Least Pattern Tree. A Significant Least Pattern Tree (SLP-Tree) is a compressed representation of significant least itemsets. There are three steps involved in constructing SLP-Tree. In the first step, the algorithm scans all transactions to determine a list of least items, LItems and frequent items, FItems (least frequent item, LFItems). In the second step, all transactions are sorted in descending order and mapping against the LFItems. It is a must in the transactions to consist at least one of the least items. In the final step, a transaction is transformed into a new path or mapped into the existing path.

***Generate Significant Least Pattern Growth* (S*LP-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. The algorithm will extract the prefix path sub-trees ending with any least item. In each of prefix path sub-tree, the algorithm will recursively execute to extract all frequent itemsets and finally built a conditional SLP-Tree.

4.2 Value Assignment

Apply Correlation and Critical Relative Support Measures. The values of association rule are derived based on Lift [26] and CRS [12, 36] measures. The processes of generating the values of association rule are taken place after all patterns and association rules are completely produced.

Discover Interesting Association Rule. From the list of valued association rules, the algorithm will begin to scan all of them. However, only those valued association rule with the correlation value that more than one and with the certain CRS value are captured and considered as Interesting ARs.

5 Result and Discussion

In order to capture the interesting ARs, the experiment employed SLP-Growth method and conducted on Intel® Core™ 2 Quad CPU at 2.33GHz speed with 4GB main memory, running on Microsoft Windows Vista. The algorithm has been developed using C# as a programming language.

The data was obtained from Division of Academic, Universiti Malaysia Terengganu in a text file and Microsoft excel format. 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. In addition, SLP-Growth algorithm with lift measurement to determine the degree of correlation of association rules was employed. There are 3,768 ARs are successfully extracted from the dataset. ARs are formed by applying the relationship of an item or many items to an item (cardinality: many-to-one). Fig. 1 depicts the correlation's classification of interesting ARs. The rule is categorized as significant and interesting if it has positive correlation, confidence is 100% and CRS value is equal to 1.0. Fig. 2 depicts the correlation of interesting ARs based on several ISupp. The result indicates that CRS successfully in producing the less number of ARs as compared to the others measures. The typical support or confidence measure alone is not a suitable measure to be employed to discover the interesting ARs. Although, correlation measure can be used to capture the interesting ARs, it ratio is still nearly 12 times larger than CRS measure. Therefore, CRS is proven to be more efficient and outperformed the benchmarked measures for discovering the interesting ARs from the dataset. Generally, the total numbers of ARs are kept decreased when the predefined Interval Supports thresholds are increased.

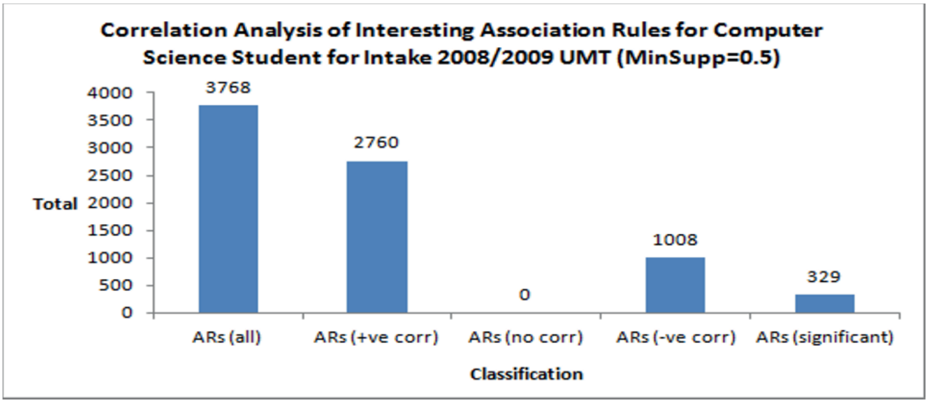
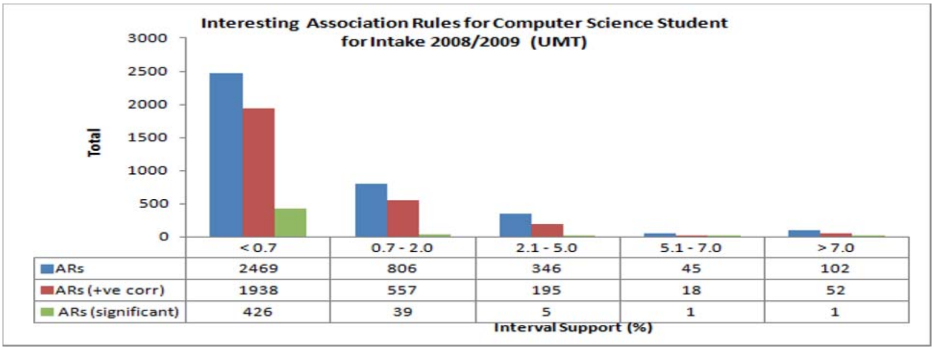


Fig. 1. Classification of ARs using correlation analysis. Only 8.73% from the total of 3,768 ARs are classified as interesting ARs



	< 0.7	0.7 - 2.0	2.1 - 5.0	5.1 - 7.0	> 7.0
🟦 ARs	2469	806	346	45	102
🟥 ARs (+ve corr)	1938	557	195	18	52
🟩 ARs (significant)	426	39	5	1	1

Fig. 2. Correlation analysis of interesting ARs using variety Interval Supports

6 Conclusion

Recently, there is an increasing interest in data mining and educational systems, making educational data mining as a new growing research community [4]. In this paper, we applied Significant Least Pattern Growth algorithm (SLP-Growth) and Critical Relative Support measure (CRS) proposed by [11,36] to mining the interesting association rules from student enrollment admission dataset. The dataset was taken from Division of Academic, Universiti Malaysia Terengganu (UMT) for the intake student 2008/2009. The results show that the interesting ARs can be extracted with the lesser number as compared to the common measures. Moreover, the results also can be analyzed by educators or the university' higher authority personnel in offering more appropriate programs to the prospect students rather than by chance.

Acknowledgment. The authors would like to thanks Universiti Malaysia Terengganu for supporting this work. The work of Tutut Herawan is supported by Excellent Research Grant Scheme no vote O7/UTY-R/SK/0/X/2013 from Universitas Teknologi Yogyakarta, Indonesia.

References

1. Yukselture, E.,and Inan, F.A.: Examining the Factors Affecting Student Dropout in an Online Certificate Program. Turkish Online Journal of Distance Education, 7(3), July 2006, 76-88 (2006)
2. Mohammad, S., and El-Masri, A.Z.: Factors Affecting Dropouts Students in Arab Open University - Bahrain Branch. International Journal of Science and Technology, 2(7), July 2012, 435-442 (2012)
3. Stearns, E., and Glennie, E.J.: When and Why Dropouts Leave High School. Youth Society. SAGE Journal, vol. 38, no. 1, September 2006, 29-57 (2006)
4. Romero, C., and Ventura, S.: Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications, 33, 135–146 (2007)
5. Ceglar, A., Roddick, J.F.: Association Mining. ACM Computing Surveys, 38(2), 1–42 (2006)
6. Agrawal, R., Imielinski, T., and Swami, A.: Database Mining: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering, 5 (6), 914–925 (1993)
7. Agrawal, R., Imielinski, T., and Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the ACM SIGMOD '93 Intl. Conf. on the Management of Data, pp. 207–216 (1993)
8. Agrawal, R., and Srikant, R.: Fast Algorithms for Mining Association Rules. In Proceedings of the 20th Intl. Conf. on Very Large Data Bases (VLDB) 1994, pp.487–499 (1994)
9. Yun, H., Ha, D., Hwang, B., and Ryu, K.H.: Mining Association Rules on Significant Rare Data Using Relative Support. The Journal of Systems and Software, 67 (3), 181-19 (2003)
10. Xiong, H., Shekhar, S., Tan, P.N., and Kumar, V.: Exploiting A Support-Based Upper Bond Pearson's Correlation Coefficient for Efficiently Identifying Strongly Correlated Pairs. In The Proceeding of ACM SIGKDD 2004, pp. 334-343 (2004)
11. Abdullah, Z., Herawan, T., and Deris, M.M.: Mining Significant Least Association Rules using Fast SLP-Growth Algorithm. In T.H. Kim and H. Adeli (Eds.): AST/UCMA/ISA/ACN

2010, LNCS, 6059, pp. 324–336. Springer Heidelberg (2010)

12. Abdullah, Z., Herawan, T., and Deris, M.M.: Scalable Model for Mining Critical Least Association Rules. In Rongbo Zhu et al. ICICA 2010, LNCS, 6377, pp. 509-516. Springer Heidelberg (2010)
13. Abdullah, Z., Herawan, T., Noraziah, A., and Deris, M.M.: Extracting Highly Positive Association Rules from Students' Enrollment Data. Procedia Social and Behavioral Sciences, 28, 107–111 (2011)
14. Abdullah, Z., Herawan, T., Noraziah, A., and Deris, M.M.: Mining Significant Association Rules from Educational Data using Critical Relative Support Approach. Procedia Social and Behavioral Sciences, 28, 97–101 (2011)
15. Abdullah, Z., Herawan, T., and Deris, M.M.: An Alternative Measure for Mining Weighted Least Association Rule and Its Framework. In J.M. Zain et al. (Eds.): ICSECS 2011, CCIS, vol. 188, II, pp. 475–485. Springer Heidelberg (2011)
16. Abdullah, Z., Herawan, T., and Deris, M.M.: Visualizing the Construction of Incremental Disorder Trie Itemset Data Structure (DOSTrieIT) for Frequent Pattern Tree (FP-Tree). In H.B. Zaman et al. (Eds.): IVIC 2011, LNCS, 7066, pp. 183–195. Springer Heidelberg (2011)
17. Herawan, T., Vitasari, P., and Abdullah, Z.: Mining Interesting Association Rules of Student Suffering Mathematics Anxiety. In J.M. Zain et al. (Eds.): ICSECS 2011, CCIS, 188, II, pp. 495–508. Springer Heidelberg (2011)
18. Abdullah, Z., Herawan, T., and Deris, M.M.: Efficient and Scalable Model for Mining Critical Least Association Rules. In a special issue from AST/UCMA/ISA/ACN 2010, Journal of The Chinese Institute of Engineer, Taylor and Francis, 35, No. 4, 27 June 2012, 547–554 (2012)
19. Herawan, T., Abdullah, Z., Noraziah, A., Deris, M.M., and Abawajy, J.H.: IPMA: Indirect Patterns Mining Algorithm. In N.T. Nguyen et al. (Eds.): ICCCI 2012, AMCCISCI, vol. 457, pp. 187–196. Springer Heidelberg (2012).
20. Herawan, T., Abdullah, Z., Noraziah, A., Deris, M.M., and Abawajy, J.H.: EFP-M2: Efficient Model for Mining Frequent Patterns in Transactional Database. In N.T. Nguyen et al. (Eds.): ICCCI 2012, LNCS, pp. 7654, 29–38. Springer Heidelberg (2012).
21. N. Ahmad, Z. Abdullah, T. Herawan, and M.M. Deris. Scalable Technique to Discover Items Support from Trie Data Structure. In B. Liu et al. (Eds.): ICICA 2012, LNCS, vol. 7473, pp. 500–507. Springer Heidelberg (2012)
22. N. Ahmad, Z. Abdullah, T. Herawan, and M.M. Deris.: WLAR-Viz: Weighted Least Association Rules Visualization. In B. Liu et al. (Eds.): ICICA 2012, LNCS, 7473, pp. 592–600. Springer Heidelberg (2012)
23. Herawan, T., and Abdullah, Z.: CNAR-M: A Model for Mining Critical Negative Association Rules. In Zhihua Cai et al. (Eds): ISICA 2012, CCIS, 316, pp. 170–179. Springer Heidelberg (2012)
24. Abdullah, Z., Herawan, T., Ahmad, N., and Deris, M.M.: DFP-Growth: An Efficient Algorithm for Mining Pattern in Dynamic Database. In B. Liu et al. (Eds.): ICICA 2012, LNCS, vol. 7473, pp. 51–59. Springer Heidelberg (2012)
25. Abdullah, Z., Herawan, T., and Deris, M.M.: Detecting Critical Least Association Rules in Medical Databasess. International Journal of Modern Physics: Conference Series, World Scientific, 9, 464–479 (2012)
26. Herawan, T., Abdullah, Z., Noraziah, A., Deris, M.M., and Abawajy, J.H.: EFP-M2: Efficient Model for Mining Frequent Patterns in Transactional Database. In N.T. Nguyen et al. (Eds.): ICCCI 2012, LNCS, 7654, pp. 29–38. Springer Heidelberg (2012)
27. Zhou, L., and Yau, S.: Association Rule and Quantitative Association Rule Mining Among Infrequent Items. Rare Association Rule Mining and Knowledge Discovery, pp.15-32. IGI-Global (2010)

28. J. Ding.: Efficient Association Rule Mining among Infrequent Items. Ph.D Thesis, University of Illinois at Chicago. (2005)
29. Y.S. Koh and N. Rountree.: Finding Sporadic Rules using Apriori-Inverse. LNCS, vol. 3518, pp. 97–106. Springer Heidelberg (2005)
30. Liu, B., Hsu, W., and Ma, Y.: Mining Association Rules with Multiple Minimum Supports, SIGKDD Explorations, pp. 337 – 341 (1999)
31. Brin, S., Motwani, R., and Silverstein, C.: Beyond Market Baskets: Generalizing ARs to Correlations. Special Interest Group on Management of Data (SIGMOD'97), pp. 265–276 (1997)
32. Omniecinski, E.: Alternative Interest Measures for Mining Associations. IEEE Trans. Knowledge and Data Engineering, 15, 57–69 (2003)
33. Y.-K. Lee, W.-Y. Kim, Y.D. Cai, J. Han. CoMine: Efficient Mining of Correlated Patterns. The Proceeding of ICDM'03 (2003)
34. Han, J., Pei, J., Yin, Y., and Mao, R.: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach∗. Data Mining and Knowledge Discovery, 8, 53–87 (2004)
35. Han, J., and Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd ed., (2006)
36. Abdullah, Z., Herawan, T., and Deris, M.M.: Tracing Significant Information using Critical Least Association Rules Model. International Journal of Innovative Computing and Applications, Inderscience, 5, 3-17 (2013)
37. Baker, R., and Yacef, K.: The State of Educational Data mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1(1), 3–17 (2010)
38. Kumar, V., and Chadha, A.: Mining Association Rules in Student's Assessment Data. International Journal of Computer Science Issues, 9(5), 3, September 2012, 211-216 (2012)
39. Garcia, E., Romero, C., Ventura, S., and Castro, C.: An Architecture for Making Recommendations to Courseware Authors using Association Rule Mining and Collaborative Filtering. User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 19, 99–132 (2009)
40. Tair, M.A.A., and El-Halees, AM.:. Mining Educational Data to Improve Students' Performance: A Case Study. International Journal of Information and Communication Technology Research, 2(2), 140-146 (2012)
41. Chandra, E., and Nandhini, K.: Knowledge Mining from Student Data. European Journal of Scientific Research, 47(1), 156-163 (2010)