# Mining Indirect Least Association Rule

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

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

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

**Abstract.** Indirect pattern can be considered as one of the interesting information that is hiding in transactional database. It corresponds to the property of high dependencies between two items that are rarely appeared together but indirectly occurred through another items. Therefore, we propose an algorithm for Mining Indirect Least Association Rule (MILAR) from the real dataset. MILAR is embedded with a scalable least measure called Critical Relative Support (CRS). The experimental results indicate that MILAR is capable in generating the indirect least association rules from the given dataset.

**Keywords:** Mining, indirect, least, association rule.

## 1 Introduction

Data mining is about making analysis convenient, scaling analysis algorithms to large databases and providing data owners with easy to use tools in helping the user to navigate, visualize, summarize and model the data [1]. In summary, the ultimate goal of data mining is more towards knowledge discovery. One of the important models and extensively studies in data mining is known as association rule mining.

Since it was first introduced by Agrawal *et al.* [2] in 1993, association rule mining has been extensively studied by many researchers [3-11]. The general aim of ARM is at discovering interesting relationship among a set of items that frequently occurred together in transactional database [12]. However, under this concept, infrequent or least items are automatically considered as not important and pruned out during rules generation. In certain domain applications, least items may also provide useful insight about the data such as competitive product analysis [13], text mining [14], web recommendation [15], biomedical analysis [16], etc. Indirect association rule [17] refers to a pair of items that are rarely occurred together but their existences are highly depending on the presence of mediator itemsets. It was first proposed by Tan *et al.* [13] for interpreting the value of infrequent patterns and effectively pruning out the uninteresting infrequent patterns. Recently, the problem of indirect association mining has become more and more important because of its various domain applications [17-21]. Generally, the studies on indirect association mining can be divided into two categories, either focusing on proposing more efficient mining algorithms [14,17,21] or extending the definition of indirect association for different

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_19,

domain applications [5,17,18]. The process of discovering indirect association rule is a nontrivial and usually relies more on the existing interesting measures that has been discussed in [13]. However, most of the measures are not properly evaluated in term of the least association rule. Therefore, in this paper we propose Mining Indirect Least Association Rule (MILAR) algorithm by utilizing the strength of Least Pattern Tree (LP-Tree) data structure [11]. In addition, Critical Relative Support (CRS) measure [22,27-35] is also embedded in the algorithm to mine the indirect least association rules among the least rules.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 explains the proposed method. This is followed by performance analysis through two experiment tests in section 4. Finally, conclusion is reported in section 5.

## 2 Related Work

Indirect association is closely related to negative association. It deals with itemsets that do not have a sufficiently highest support. The negative associations' rule was first pointed out by Brin *et al.* [23]. The focused on mining negative associations is better that on finding the itemsets that have a very low probability of occurring together. Indirect associations provide an effective way to detect interesting negative associations by discovering only frequent itempairs that are highly expected to be frequent.

Until this recent, the important of indirect association between items has been discussed in many literatures. Tan *et al.* [13] proposed INDIRECT algorithm to extract indirect association between itempairs using the famous Apriori technique. Wan *et al.* [14] introduced HI-Mine algorithm to mine a complete set of indirect associations. HI-Mine generates indirect itempair set (IIS) and mediator support set (MSS), by recursively building the HI-struct from database. IS measure [24] is used as a dependence measure. Lin *et al.* [25] suggested GIAMS as an algorithm to mine indirect associations over data streams rather than static database environment. GIAMS contains two concurrent processes called PA-Monitoring and IA-Generation. In term of dependence measure, IS measure [24] is again adopted in the algorithm. Chen *et al*. [17] proposed an indirect association algorithm that was similar to HI-mine, namely MG-Growth. In this algorithm, temporal support and temporal dependence are used in this algorithm. Kazienko [15] presented IDARM* algorithm to extracts complete indirect associations rules. The main idea of IDARM* is to capture the transitive page from user-session as part of web recommendation system. A simple measure called Confidence [2] is employed as dependence measure. Lin *et al*. [36] presented EMIA-LM algorithm for mining indirect association rules over web data stream. The preliminary experiments also showed that EMIA-LM is better than HI-mine* for static data in term of computational speed and memory consumption. Liu *et al*. [37] suggested FIARM (Filtering-Based Indirect Association Rule Mining) algorithm to analyze gene microarray data. It is Apriori-based algorithm. In the analysis, Gene Ontology is employed to verify the accuracy of the relationships.

# 3 The Proposed Method

## 3.1 Association Rule

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 $\text{supp}(X)$ 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 $X\cap Y=\phi$. The sets X and Y are called antecedent and consequent, respectively.*

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

**Definition 5.** *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 $X\cup Y$ to the number of transactions in D contain X. Thus*

$$\text{conf}(X\Rightarrow Y)=\frac{\text{supp}(X\Rightarrow Y)}{\text{supp}(X)}$$

**Definition 6**. (Least Items). *An itemset X is called least item if $\text{supp}(X)<\alpha$, where $\alpha$ is the minimum support (minsupp)*

The set of least item will be denoted as Least Items and

$$\text{Least Items}=\{X\subset I\mid\text{supp}(X)<\alpha\}$$

**Definition 7**. (Frequent Items). *An itemset X is called frequent item if $\text{supp}(X)\ge\alpha$, where $\alpha$ is the minimum support.*

The set of frequent item will be denoted as Frequent Items and

$$\text{Frequent Items}=\{X\subset I\mid\text{supp}(X)\ge\alpha\}$$

### 3.2 Indirect Association Rule

**Definition 8.** *An itempair* $\{X\,Y\}$ *is indirectly associated via a mediator M, if the following conditions are fulfilled:*

1. $\text{supp}(\{X,Y\}) < \text{t}_\text{s}$ (itempair support condition)
2. There exists a non-empty set *M* such that:
   a. $\text{supp}(\{X\} \cup M) \geq t_m$ and $\text{supp}(\{Y\} \cup M) \geq t_m$ (mediator support condition)
   b. $\text{dep}(\{X\}, M) \geq t_d$ and $\text{dep}(\{Y\}, M) \geq t_d$, where $\text{dep}(A, M)$ is a measure of dependence between itemset *A* and *M* (mediator dependence measure)

The user-defined thresholds above are known as itempair support threshold $(t_s)$, mediator support threshold $(t_m)$ and mediator dependence threshold $(t_d)$, respectively. The itempair support threshold is equivalent to *minsupp* $(\alpha)$. Normally, the mediator support condition is set to equal or more than the itempair support condition $(t_m \geq t_s)$

The first condition is to ensure that $(X,Y)$ is rarely occurred together and also known as least or infrequent items. In the second condition, the first-sub-condition is to capture the mediator *M* and for the second-sub-condition is to make sure that *X* and *Y* are highly dependence to form a set of mediator.

**Definition 9**. (Critical Relative Support). *A Critical Relative Support (CRS) is a formulation of maximizing relative frequency between itemset and their Jaccard similarity coefficient.*

The value of Critical Relative Support denoted as CRS and

$$\text{CRS}(A,B) = \max\left(\left(\frac{\text{supp(A)}}{\text{supp(B)}}\right),\left(\frac{\text{supp(B)}}{\text{supp(A)}}\right)\right) \times \left(\frac{\text{supp}(A \Rightarrow B)}{\text{supp(A)} + \text{supp(B)} - \text{supp}(A \Rightarrow B)}\right)$$

CRS value is between 0 and 1, and is determined by multiplying the highest value either supports of antecedent divide by consequence or in another way around with their Jaccard similarity coefficient. It is a measurement to show the level of CRS between combination of the both Least Items and Frequent Items either as antecedent or consequence, respectively. Here, Critical Relative Support (CRS) is employed as a dependence measure for 2(a) in order to mine the desired Indirect Association Rule.

### 3.2 Algorithm Development

**Determine Minimum Support.** Let *I* is a non-empty set such that $I = \{i_1, i_2, \cdots, i_n\}$, and *D* is a database of transactions where each *T* is a set of items such that $T \subset I$. An itemset is a set of item. A *k*-itemset is an itemset that contains *k* items. From Definition 6, an itemset is said to be least (infrequent) if it has a support count less than $\alpha$.

**Construct LP-Tree.** A Least Pattern Tree (LP-Tree) is a compressed representation of the least itemset. It is constructed by scanning the dataset of single transaction at a

time and then mapping onto a new or existing path in the LP-Tree. Items that satisfy the $\alpha$ (Definition 6 and 7) are only captured and used in constructing the LP-Tree.

**Mining LP-Tree.** After the LP-Tree is fully constructed, the mining process will begin by implementing the bottom-up strategy. Hybrid 'Divide and conquer' method is employed to decompose the tasks of mining desired pattern. LP-Tree utilizes the strength of hash-based method during constructing itemset in support descending order.

**Construct Indirect Patterns.** The pattern is classified as indirect association pattern if it fulfilled with the two conditions. The first condition is elaborated in Definition 8 where it contains three sub-conditions. One of them is mediator dependence measure. CRS from Definition 9 is employed as mediator dependence measure between itemset in discovering the indirect patterns. Fig. 1 shows a complete graphical representation of Mining Indirect Least Association Rule Algorithm (MILAR).

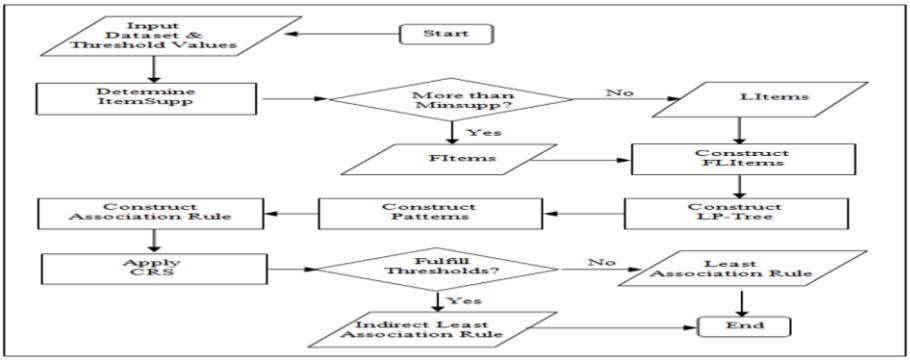


**Fig. 1.** A Complete Graphical Representation of MILAR Algorithm

## 4 Experiment Test

In this section, the analysis is made by comparing the total number of association rules being extracted based on the predefined thresholds using our proposed algorithm, MILAR. Here, three items are involved in forming a complete association rule; two items as an antecedent and one item as a consequence. The mediator is appeared as a part of antecedent. The experiment has been performed on Intel® Core™ 2 Quad CPU at 2.33GHz speed with 4GB main memory, running on Microsoft Windows Vista. All algorithms have been developed using C# as a programming language.

The dataset used in the experiment is a language anxiety dataset. It was taken from a survey on exploring language anxiety among engineering students at Universiti Malaysia Pahang (UMP) [3]. The respondents were 770 students, consisting of 394 males and 376 females. They are undergraduate students from five engineering based faculties, i.e., 216 students from Faculty of Chemical and Natural Resources Engineering (FCNRE), 105 students from Faculty of Electrical and Electronic Engineering (FEEE), 226 students from Faculty of Mechanical Engineering (FME),

178 students from Faculty of Civil Engineering and Earth Resources (FCEER), and 45 students from Faculty of Manufacturing Engineering and Technology Management (FMETM). To this, we have a dataset comprises the number of transactions (student) is 770 and the number of items (attributes) is 5.

Different Interval Supports were employed in the experiment. Fig. 2 shows the performance analysis against the dataset. Minimum Support (minsupp or $\alpha$) and Mediator Support Threshold $(t_m)$ are set to 30% and 10%, respectively. Varieties of minimum CRS (min-CRS) were employed in the experiment. During the performance analysis, 286 least association rules and 152 indirect least association rules were produced, respectively. The general trend was, the total number of indirect least association rules were kept reducing when the values of min-CRS were kept increasing. However, there are no changes in term of total least association rules and indirect least association rules when the min-CRS values were in the range of 0.15 until 0.20.

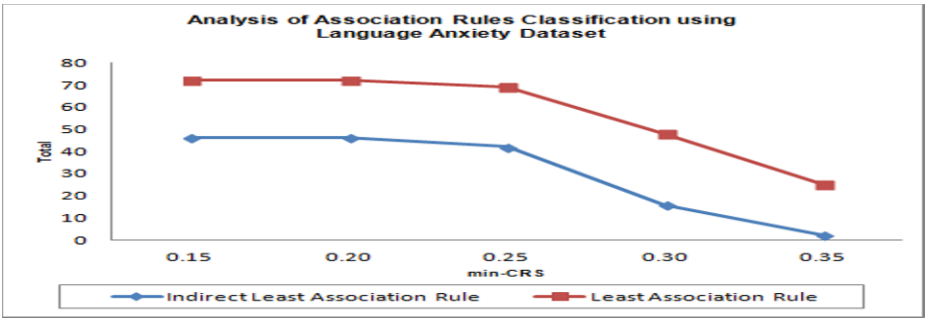


**Fig. 2.** Analysis of the Generated Association Rules against Language Anxiety Dataset with variation of min-CRS

## 5 Conclusion

Mining indirect least association rules from data repository is a very useful and nontrivial study in dealing with the rarity cases. In fact, it may contribute into discovering of a new knowledge which cannot be obtained through typical association rules approach. Therefore, we proposed Mining Indirect Least Association Rule (MILAR) algorithm to extract the hidden indirect least association rules from the data repository. MILAR algorithm embeds with a scalable measure called Critical Relative Support (CRS) rather than the common interesting measures in data mining. We conducted the experiment based on a real dataset. The result shows that MILAR algorithm can successfully generate the different number of indirect least association rules based on the variety of threshold values. It is also expected that the obtained information can provide a new insight for subject-matter experts to do further investigation.

## References


1. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P.: From Data Mining to Knowledge Discovery in Databases. Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, Menlo Park, USA (1996)
2. Agrawal, R., & Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases.: In Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487-499 (1994)
3. Mannila, H., Toivonen, H., and Verkamo, A.I.: Discovery of Frequent Episodes in Event Sequences. Data Mining and Knowledge Discovery, 1, 259–289 (1997)
4. Park, J.S., Chen, M.S., and Yu, P.S.: An Effective Hash-based Algorithm for Mining Association Rules. In: Proceedings of the ACM-SIGMOD Intl. Conf. Management of Data (SIGMOD'95), pp. 175–186. ACM Press (1995)
5. Savasere, A., Omiecinski, E., and Navathe, S.: An Efficient Algorithm for Mining Association Rules in Large Databases. In: Proceedings of the 21st Intl. Conf. on Very Large Data Bases (VLDB'95), pp. 432–443. ACM Press (1995)
6. Fayyad, U., Patesesky-Shapiro, G., Smyth, P. and Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. MIT Press, MA (1996)
7. Bayardo, R.J.: Efficiently Mining Long Patterns from Databases. In: Proceedings of the ACM-SIGMOD International Conference on Management of Data (SIGMOD'98), pp. 85–93. ACM Press (1998)
8. Zaki, M.J. and Hsiao, C.J.: CHARM: An Efficient Algorithm for Closed Itemset Mining. In: Proceedings of the 2002 SIAM Intl. Conf. Data Mining, pp. 457–473. SIAM (2002).
9. Agarwal, R., Aggarwal, C., and Prasad, V.V.V.:A Tree Projection Algorithm for generation of Frequent Itemsets. Journal of Parallel and Distributed Computing 61, 350–371 (2001)
10. Liu, B., Hsu, W. and Ma, Y.: Mining Association Rules with Multiple Minimum Support. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.337 – 341. ACM Press (1999)
11. 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)
12. Leung, C.W., Chan S.C and Chung, F.: An Empirical Study of a Cross-level Association Rule Mining Approach to Cold-start Recommendations. Knowledge-Based Systems, 21(7), October 2008, 515–529 (2008)
13. Tan, P.N., Kumar, V. and Srivastava, J.: Indirect Association: Mining Higher Order Dependences in Data. In: Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 632-637. Springer Heidelberg (2000)
14. Wan, Q. and An, A.: An Efficient Approach to Mining Indirect Associations. Journal Intelligent Information Systems, 27(2), 135–158 (2006)
15. Kazienko, P.: Mining Indirect Association Rules for Web Recommendation. International Journal of Applied Mathematics and Computer Science, 19(1), 165–186 (2009)
16. Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J. and Ananiadou, S.: Discovering and Visualizing Indirect Associations between Biomedical Concepts. Bioinformatics, 27(13), 111-119 (2011).
17. Chen, L., Bhowmick, S.S. and Li, J.: Mining Temporal Indirect Associations. In: PAKDD2006, LNAI 3918, pp. 425-434. Springer Heidelberg (2006)
18. Cornelis, C., Yan, P., Zhang, X. and Chen, G.: Mining Positive and Negative Association from Large Databases. In: Proceedings of the 2006 IEEE International Conference on Cybernatics and Intelligent systems, pp. 1-6. IEEE (2006)
19. Kazienko, P. and Kuzminska, K.: The Influence of Indirect Association Rules on Recommendation Ranking Lists. In: Proceeding of the 5th Intl. Conf. on Intelligent Systems Design and Applications, pp. 482-487 (2005)

20. Tseng, V.s, Liu, Y.C and Shin J.W.: Mining Gene Expression Data with Indirect Association Rules. In Proceeding of the 2007 National Computer Symposium (2007)
21. Wu, X., Zhang, C. and Zhang, S. Efficient Mining of Positive and Negative Association Rules. ACM Transaction on Information Systems, 22(3), 381-405 (2004)
22. 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, 2011, 28, 97 – 191. Elsevier (2011)
23. Brin, S., Motwani, R., Ullman, J., and Tsur, S.: Dynamic Itemset Counting and Implication Rules for Market Basket Data. In: Proceedings of the International ACM SIGMOD Conference, pp. 255–264. ACM Press (1997)
24. Tan, P., Kumar, V., and Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns. In: Proceedings of the 8th Intl. Conf. on Knowledge Discovery and Data Mining, pp.32-41. (2002)
25. Lin, W-Y., Wei, Y-E., and Chen, C-H. Generic Approach for Mining Indirect Association Rules in Data Streams. LNCS 6703, pp. 95-104. Springer Heidelberg. (2011).
26. UCI Machine Learning Repository, http:// archive.ics.uci.edu/ml/datasets/
27. Herawan, T., Vitasari, P., and Abdullah, Z.: Mining Interesting Association Rules of Students Suffering Study Anxieties Using SLP-Growth Algorithm. International Journal of Knowledge and Systems Sciences, 3(2), 24-41 (2012)
28. Abdullah, Z., Herawan, T. Noraziah, A. 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)
29. 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)
30. Abdullah, Z., Herawan, T. Noraziah, A. and Deris, M.M and Abawajy, J.H..: IPMA: Indirect Patterns Mining Algorithm. In N.T. Nguyen et al. (Eds.): ICCCI 2012, AMCCISC, 457, pp. 187–196. Springer Heidelberg (2012)
31. Herawan, T., P. Vitasari, and Abdullah, Z.: Mining Interesting Association Rules of Student Suffering Mathematics Anxiety. In J.M. Zain et al. (Eds.): ICSECS 2011, CCIS, vol. 188, II, pp. 495–508. Springer Heidelberg (2011)
32. 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)
33. 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)
34. 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)
35. 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)
36. Lin, W-Y., and Chen, Y-C.: A Mediator Exploiting Approach for Mining Indirect Associations from Web Data Streams. IEEEXplore in the 2nd Intl. Conf. on Innovations in Bio-inspired Computing and Applications (IBICA) 2011, pp. 183-186 (2011)
37. Liu, Y-C., Shin, J.W., and Tseng, V.S.: Discovering Indirect Gene Associations by Filtering-Based Indirect Association Rule Mining, International Journal of Innovative Computing, Information and Control, 7(10), 6041–6053 (2011)