# Origin based Association Rule Mining using multiple MASP tree[☆]

Elsevier[1]

*Radarweg 29, Amsterdam*

*Elsevier Inc[a,b], Global Customer Service[b,*]*

[a] *1600 John F Kennedy Boulevard, Philadelphia*
[b] *360 Park Avenue South, New York*

**Abstract**

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.

*Keywords:* data-mining, Association Rule Mining, frequent-itemset mining

## 1. Introduction

Association rule mining is a rule-based machine learning procedure to find interesting patterns in the transaction database based on individual and conditional frequencies. In the traditional approach, two steps are involved in generating rules. First, generate all frequent itemsets and pruned non-frequent ones and then in the second stage rules are derived from those frequent itemsets. An association rule e.g. {bread, milk} ⇒ {butter} in market basket analysis means if one purchase bread and milk together it is highly likely that they will also buy butter. Apart from market basket analysis, association rule mining is useful in intrusion detection, bioinformatics, and many other applications.

In 2014 O. M. Soyal [1] proposed a new approach to extract mostly associated sequential patterns (MASPs) using less computational resources in terms of time

[☆] Fully documented templates are available in the elsarticle package on CTAN.

[*] Corresponding author
*Email address:* support@elsevier.com (Global Customer Service)
*URL:* www.elsevier.com (Elsevier Inc)

[1] Since 1880.

and memory while generating a long sequence of patterns that have the highest co-occurrence.

This approach may produce different outcomes if we change the order of items in transactions. We propose an approach which is order independent. An association rule of the form A $\Rightarrow$ B must satisfy the threshold support and threshold confidence i.e. probability of occurrence of A and B together must surpass threshold support, and the probability of occurrence of B in transactions containing A must be greater than or equal to threshold confidence. It means, to calculate support and confidence, it is required to traverse complete transaction database. To generate all rules containing a particular item x it is reasonable to ignore all transactions(for calculating support and confidence) that come before the transaction in which that particular item appears for the first time. Embedding these two changes to the Omer M. Soyal [1] approach is the basis of our research.

## 2. Related works

In 1994 R. Agrawal, et al. published non-trivial algorithm(Apriori) [2] for finding association rules in large databases of the sales transaction. Apriori algorithm produces association rules in two steps. First generates all frequent itemsets(prune non-frequent candidate itemsets) and then make rules from those itemsets. This algorithm first finds frequent itemsets of length one then frequent itemsets of length 2 using frequent itemsets of length 1 and so on until generation of all frequent itemsets. This algorithm gave the better result than the previously known fundamental algorithms AIS [3], SETM [4]. In 1996 Fukuda, et al. [5] proposed an approach to find two-dimensional association rules. A state in this scenario is of the form $((X, Y) \in P) \Rightarrow (Z = z)$ where X and Y are numeric attributes, P is a subspace of 2-D plane, and Z is boolean attribute i.e. z can be either true or false. E.g. $(\text{Age} \in [30, 50] \wedge \text{Balance} \in [10^5, 10^6]) \Rightarrow (\text{CardLoan} = \text{yes})$. It means if a bank user age and balance lies in the given subspace it is very likely that they will use card loan. This approach

works for specific types of structured data. R. Feldman, et al.(1997) [6] introduced the notion of maximal association rules. These are the rules extracted from frequent maximal itemsets. Frequent maximal itemsets are those itemsets which appear just once among all transactions. It is useful in finding association rules containing negated attributes. As an example a rule {milk, ¬bread} ⇒ {¬butter} contains negated attributes. It means if a user purchases milk but not bread then the probability that the user will not buy butter is very high. This approach helps to capture inference rules which might be lost using regular associations. Till now items in transaction databases were treated uniformly. In 1998 C.H. Cai, et al. [7] gave an approach to find association rules which take into account weight(importance) of items in transaction databases. FP-Growth algorithm(2000) [8] also take two steps. The second phase is same as apriori. FP-Growth does not generate candidate frequent itemsets. First, it creates a tree(FP-Tree) and then finds frequent itemsets. This algorithm is about an order of magnitude faster than the Apriori algorithm. Lin, Weiyang, et al. [9] proposed an approach that uses association rule mining for collaborative recommender systems. This approach does not require threshold support value. Instead, based on the number of rules(given) to be generated, threshold support is decided by the system. Thus it reduced the running time and produced enough rules for good recommendation performance. In 2004 F. Conen, et al. [10] proposed two structures(T-Trees and P-Trees) which offer improvement concerning storage and execution time. In 2005 K. G. Srinivasa, et al. [11] took advantage of genetic algorithms principles to generate large itemsets within dynamic transaction database. Their algorithm was better than the pre-existing FUP and E-Apriori in terms of execution time and scalability. If transaction database is static then life will be easy. In other scenario transaction database keeps on changing at high speed leading to change in data distribution. Hence it will be difficult to apply previously mentioned Association Rule Mining techniques. Jiang, et al.(2006) [12] came up with an approach to overcome this difficulty.

## References


[1] O. M. Soysal, Association rule mining with mostly associated sequential patterns, Expert Systems with Applications 42 (5) (2015) 2582 – 2592. doi:http://dx.doi.org/10.1016/j.eswa.2014.10.049.

[2] R. Agrawal, R. Srikant, Fast algorithms for mining association rules (1994) 487–499.

[3] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, SIGMOD Rec. 22 (2) (1993) 207–216. doi:10.1145/170036.170072.
URL http://doi.acm.org/10.1145/170036.170072

[4] M. Houtsma, A. Swami, Set-oriented mining of association rules, Research Report RJ 9567, IBM Almaden Research Center, San Jose, California.

[5] T. Fukuda, Y. Morimoto, S. Morishita, T. Tokuyama, Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization, SIGMOD Rec. 25 (2) (1996) 13–23. doi:10.1145/235968.233313.
URL http://doi.acm.org/10.1145/235968.233313

[6] R. Feldman, Y. Aumann, A. Amir, A. Zilberstein, W. Kloesgen, Maximal association rules: a new tool for mining for keyword cooccurrences in document collections.

[7] C. H. Cai, A. W. C. Fu, C. H. Cheng, W. W. Kwong, Mining association rules with weighted items, in: Database Engineering and Applications Symposium, 1998. Proceedings. IDEAS'98. International, 1998, pp. 68–77. doi:10.1109/IDEAS.1998.694360.

[8] J. Han, J. Pei, Y. Yin, R. Mao, Mining frequent patterns without candidate generation: A frequent-pattern tree approach, Data Mining and Knowledge Discovery 8 (1) (2004) 53–87. doi:10.1023/B:DAMI.0000005258.31418.

83.
URL http://dx.doi.org/10.1023/B:DAMI.0000005258.31418.83

[9] W. Lin, S. A. Alvarez, C. Ruiz, Efficient adaptive-support association rule mining for recommender systems, Data Mining and Knowledge Discovery 6 (1) (2002) 83–105. doi:10.1023/A:1013284820704.
URL http://dx.doi.org/10.1023/A:1013284820704

[10] F. Coenen, P. Leng, S. Ahmed, Data structure for association rule mining: T-trees and p-trees, IEEE Transactions on Knowledge and Data Engineering 16 (6) (2004) 774–778. doi:10.1109/TKDE.2004.8.

[11] P. D. Shenoy, K. G. Srinivasa, K. R. Venugopal, L. M. Patnaik, Dynamic association rule mining using genetic algorithms, Intell. Data Anal. 9 (5) (2005) 439–453.
URL http://dl.acm.org/citation.cfm?id=1239098.1239101

[12] N. Jiang, L. Gruenwald, Research issues in data stream association rule mining, SIGMOD Rec. 35 (1) (2006) 14–19. doi:10.1145/1121995.1121998.
URL http://doi.acm.org/10.1145/1121995.1121998