# Association Rule Mining and its Application to MPIS

Raymond Chi-Wing Wong, Ada Wai-Chee Fu
The Chinese University of Hong Kong

## INTRODUCTION

Association rule mining (Agrawal, Imilienski and Swami, 1993) has been proposed for understanding the relationships among items in transactions or market baskets. For instance, if a customer buys butter, what is the chance that he/she buys bread at the same time? Such information may be useful for decision makers to determine strategies in a store.

More formally, given a set $I = \{I_1, I_2, \ldots, I_n\}$ of *items* (e.g. carrot, orange and knife, in a supermarket). The database contains a number of transactions. Each transaction $t$ is a binary vector with $t[k]=1$ if $t$ bought item $I_k$ and $t[k]=0$ otherwise(e.g. {1, 0, 0, 1, 0}). An association rule is of the form $X \rightarrow I_j$, where $X$ is a set of some items in $I$, and $I_j$ is a single item not in $X$ (e.g. {Orange, Knife} ➔ Plate).

A transaction $t$ satisfies $X$ if for all items $I_k$ in $X$, $t[k] = 1$. The *support* for a rule $X \rightarrow I_j$ is the fraction of transactions that satisfy the union of $X$ and $I_j$. A rule $X \rightarrow I_j$ has *confidence* c% if and only if c% of transactions that satisfy $X$ also satisfy $I_j$.

The mining process of association rule can be divided into two steps.

1. **Frequent Itemset Generation:** generate all sets of items that have support greater than a certain threshold, called *minsupport*
2. **Association Rule Generation:** from the frequent itemsets, generate all association rules that have confidence greater than a certain threshold called *minconfidence*

Step 1 is much more difficult compared with Step 2. Thus, researchers (Agrawal, Imilienski &Swamin, 1993; Han, Pei & Yin, 2000; Han, Wang, Lu and Tzvetkov, 2002; Liu, Pan, Wang & Han, 2002; Fu, Kwong & Tang, 2000) have focused on the studies of frequent itemset generation.

Different algorithms have been proposed for finding frequent itemsets. The ***Apriori Algorithm*** is a well-known approach which is proposed by Agrawal & Srikant (1994). It is an iterative approach and there are two steps in each iteration. The first step generates a set of candidate itemsets. Then, the second step prunes all disqualified candidates (i.e. all infrequent itemsets). The iterations begin with size 2 itemsets and the size is incremented at each iteration. The algorithm is based on the ***closure property*** of frequent itemsets: if a set of items is frequent, then all its proper subsets are also frequent. The weaknesses of this algorithm are the generation of a large number of candidate itemsets and the requirement to scan the database once in each iteration.

A data structure called ***FP-tree*** and an efficient algorithm called FP-growth are proposed by Han, Pei & Yin (2000) to overcome the above weaknesses. The idea of FP-tree is fetching all transactions from the database and inserting them into a compressed tree structure. Then, algorithm FP-growth reads from the structure FP-tree to mine frequent itemsets.

## VARIATIONS IN ASSOCIATION RULES

Many variations on the above problem formulation have been suggested. The association rules can be classified based on the following (Han and Kamber, 2000):

### 1. Association Rules based on the Type of Values of Attribute

Based on the type of values of attributes, there are two kinds – boolean association rule, which is presented above, and quantitative association rule. ***Quantitative association rule*** describes the relationships among some quantitative attributes (e.g. income and age). An example is income(40K..50K) ➔ age(40..45). One proposed method is grid-based - first dividing each attribute into a fixed number of partitions (Association Rule Clustering System (ARCS) inLent, Swami & Widom 1997). There is also a non-grid based approach which does not require any fixed number of partition initially (Srikant & Agrawal, 1996; Zhang, Padmanabhan, & Tuzhilin, 2004). Srikant & Agrawal (1996) proposes to partition quantitative attributes *dynamically* and to merge the partitions based on a measure of *partial completeness*.

### 2. Association Rules based on the Dimensionality of Data

Association rules can be divided into *single-dimensional association rules* and *multi-dimensional association rule*s One example of single-dimensional rule is buys({Orange, Knife}) ➔ buys(Plate) which contains only the dimension *buys*. Multi-dimensional association rule is the one containing attributes for more than one dimension. For example, income(40K..50K) ➔ buys(Plate). One mining approach is to borrow the concept of *data cube* in the field of data warehousing. Figure 1 shows a lattice for the data cube for the dimensions age, income and buys. Researchers (Kamber, Han & Chiang, 1997) have applied the data cube model and used the *aggregate* techniques for mining.

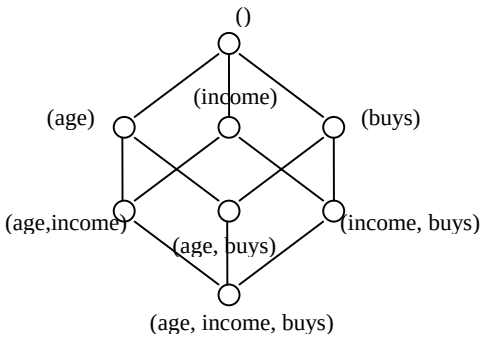


Figure 1: A lattice showing the data cube for the dimensions age, income and buys.

### 3. Association Rules based on the Level of Abstractions of Attribute

The rules discussed in previous sections can be viewed as single-level association rule. A rule which references different levels of abstraction of attributes is called a ***multilevel association rule***. Suppose there are two rules – income(10K..20K) ➔ buys(fruit) and income(10K..20K) ➔ buys(orange). There are two different levels of abstractions in these two rules because "fruit" is a higher-level abstraction of "orange". Han & Fu (1995) and Srikant & Agrawal (1995) apply a top-down strategy to the concept hierarchy in the mining of frequent itemsets.

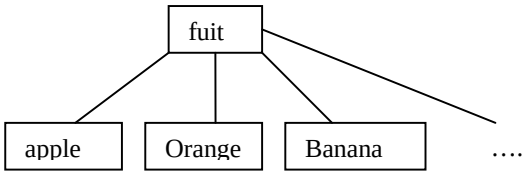


Figure 2: A concept hierarchy of the fruit

## OTHER EXTENSIONS TO ASSOCIATION RULE MINING

There are other extensions to association rule mining. Some of them (Bayardo, 1998) find ***maxpattern*** (i.e. maximal frequent patterns) while others (Pei, Han & Mao, 2000; Zaki & Hsiao, 2002) find ***frequent closed itemsets***. Maxpattern is a frequent itemset which does not have a frequent item superset. A frequent itemset is a frequent closed itemsets if there exists no itemset $X'$ such that (1) $X \subset X'$ and (2) $\forall$ transactions $t$, $X$ is in $t$ implies $X'$is in $t$. These considerations can reduce the resulting number of frequent itemsets significantly.

Another variation of the frequent-itemset problem is mining ***top-K frequent*** itemsets (Fu, Kwong & Tang, 2000; Cheung & Fu, 2004). The problem is to find $K$ frequent itemsets with the greatest supports. It is often more reasonable to assume the parameter $K$, instead of the data-distribution dependent parameter of

minsupport because the user typically would not have the knowledge of the data distribution before data mining.

The other variations of the problem are the incremental update of mining association rules (Sarda & Srinivas, 1998; Hidber, 1999), constraint-based rule mining (Ng, Lakshmanan, Han & Pang, 1998; Pei & Han, 2000; Grahne & Lakshmanan, 2000), distributed and parallel association rule mining (Agrawal & Shafer, 1996; Zaki, 1999; Schuster, Wolff & Trock, 2003; Gilburd, Schuster, & Wolff, 2004), association rule mining with multiple minimum supports/without minimum support (Liu, Hsu & Ma, 1999; Chiu, Wu, & Chen, 2004), association rule mining with weighted item and weight support (Cai, Fu, Cheng & Kwong, 1998; Tao, Murtagh & Farid, 2003), and fuzzy association rule mining (Kuok, Fu & Wong 1998).

Association rule mining has been integrated with other data mining problems. There have been integration of classification and association rule mining (Liu, Hsu & Ma, 1998) and the integration of association rule mining with relational database systems (Sarawagi, Thomas & Agrawal, 1998).

## APPLICATION OF ASSOCIATION RULES to MPIS

Other than market basket analysis (Blischok, 1995), association rules can also help in applications such as intrusion detection (Lee, Stolfo & Mok, 1999), heterogeneous genome data (Satou, Shibayama, Ono, et al, 1997), mining remotely sensed images/data (Dong, Perrizo, Ding and Zhou, 2000) and product assortment decisions (Brijs, Swinnen, Vanhoof & Wets, 1999; Brijs, Goethals, Swinnen, Vanhoof & Wets, 2000; Wang & Su, 2002; Wong, Fu & Wang, 2003; Wong & Fu, 2004). Here we focus on the application on product assortment decisions as it is one of very few examples where the association rules are not the end mining results.

Transaction database in some applications can be very large. For example Hedberg (1995) quoted that Wal-Mart kept about 20 million sales transactions per day. Such data requires sophisticated analysis. As pointed out by Blischok (1995), a major task of talented merchants is to pick the profit generating items and discard the losing items. It may be simple enough to sort items by their profit and do the selection. However, this ignores a very important aspect in market analysis - the cross-selling effect. There can be items that do not generate much profit by themselves but they are the catalysts for the sales of other profitable items. Recently, some researchers (Kleinberg, Papadimitriou & Raghavan, 1998) suggest that concepts of association rules can be used in the item selection problem with the consideration of relationships among items.

One example of the product assortment decisions is Maximal-Profit Item Selection (**MPIS**) with Cross-selling considerations (Wong, Fu & Wang, 2003). Consider the major task of merchants to pick profit generating items and discard the losing items. Assume we have a history record of the sales (transactions) of all items. This problem is to select a subset from the given set of items so that the estimated profit of the resulting selection is maximal among all choices.

Suppose a shop carries office equipments of monitors, keyboards and telephones, with profits of $1000K, $100K and $300K, respectively. If now the shop decides to remove one of the three items from its stock, the question is which two we should choose to keep. If we simply examine the profits, we may choose to keep monitors and telephones, and so the total profit is $1300K. However, we know

that there is strong cross-selling effect between monitor and keyboard (see the table below). If the shop stops carrying keyboard, the customers of monitor may choose to shop elsewhere to get both items. The profit from monitor may drop greatly, and we may be left with profit of $300K from telephones only. If we choose to keep both monitors and keyboards, then the profit can be expected to be $1100K which is higher.

| Monitor | Keyboard | Telephone |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 1 |

MPIS will give us the desired solution. MPIS utilizes the concept of the relationship between selected items and unselected items. Such relationship is modeled by the cross-selling factor. Suppose $d$ is the set of unselected items and $I$ is the selected item. A ***loss rule*** is proposed in the form $I \rightarrow \Diamond d$, where $\Diamond d$ means the purchase of any item in $d$. The rule indicates that from the history, whenever a customer buys the item $I$, he/she also buys at least one of the items in $d$. Interpreting this as a pattern of customer behavior, and assuming that the pattern will not change even when some items were removed from the stock, if none of the items in d are available then the customer also will not purchase I. This is because if the customer still purchases $I$, without purchasing any items in d, then the pattern would be changed. Therefore, the higher the confidence of $I \rightarrow \Diamond d$, the more likely the profit of $I$ should not be counted. This is the reasoning behind the above definition. In the above example, suppose we choose monitor and telephone. Then, d = {keyboard}. All profits of monitor will be lost because, in the history, we find conf($I \rightarrow \Diamond d$)=1, where $I$ = monitor. This example illustrates the importance of the consideration of cross-selling factor in the profit estimation.

Wong, Fu & Wang (2003) proposes two algorithms to deal with this problem. The first one is an optimization approach called QP. They approximate the total profit of the item selection in quadratic form and solve a quadratic optimization problem. The second one is a greedy approach called MPIS_Alg. This algorithm prune items iteratively according to an estimated function based on the formula of the total profit of the item selection until $J$ items remain.

Another the product assortment decision problem is studied in (Wong & Fu, 2004), which addresses the problem of selecting a set of marketing items in order to boost the sales of the store.

## FUTURE TRENDS

A new area for investigation of the problem of the frequent itemsets is mining ***data streaming*** for frequent itemsets (Manku & Motwani, 2002; Giannella, Han, Pei, Yan & Yu, 2003; Chang & Lee, 2003). In such kind of problem, the data is so massive that all data cannot be stored in the memory of a computer and cannot be processed by traditional algorithms. The objective of all proposed algorithm is to store as few as possible and to minimize the error generated by some estimation in the model. For instance, Yu, Chong, Lu & Zhou (2004) recently proposed a false-negative oriented algorithm for frequent itemset mining.

***Privacy preservation*** on the association rule mining is also rigorously studied in these few years (Vaidya & Clifton, 2002; Rizvi & Haritsa, 2002; Agrawal, Evfimievski & Srikant, 2003). The problem is to mine from two or more different sources without exposing individual transaction data to each others.

## CONCLUSION

Association rule mining plays an important role in the literature of data mining. It poses many challenging issues for the development of efficient and effective methods. After taking a closer look, we find that the application of association rules requires much more investigations in order to aid in more specific targets. We may see a trend towards the study of applications of association rules.

## REFERENCES


Agrawal, R., Evfimievski, A.and Srikant, R. (2003) Information Sharing Across Private Database, SIGMOD

Agrawal, R., Imilienski, T. and Swami (1993). Mining Association Rules between Sets of Items in Large Databases, *SIGMOD*

Agrawal, R. and Shafer, J. C. (1996) Parallel Mining of Association Rules, *IEEE Transactions on Knowledge and Data Engineering*

Agrawal, R., and Srikant, R. (1994). Fast Algorithms for Mining Association Rules, *VLDB*

Bayardo, R. J. (1998) Efficiently Mining Long Patterns from Databases, *SIGMOD*

Blischok, T (1995). Every transaction tells a story, *Chain Store Age Executive with Shopping Center Age* 71 (3), 50-57

Brijs, T., Goethals, B., Swinnen, G., Vanhoof, K. and Wets, G.(2000). A Data Mining Framework for Optimal Product Selection in Retail Supermarket Data: The Generalized PROFSET Mo del, *SIGKDD*

Brijs, T., Swinnen, G., Vanhoof, K. and Wets, G.(1999). Using Association Rules for Product Assortment Decisions: A Case Study, *SIGKDD*

Cai, C.H., Fu, A.W.-C., Cheng, C.H. and Kwong, W.W. (1998) Mining Association Rules with Weighted Items, *IDEAS*

Chang, J. H. and Lee, W. S. (2003). Finding Recent Frequent Itemsets Adaptively over Online Data Streams, *SIGKDD*

Cheung, Y.L. and Fu, A. W.-C. (2004), Mining Association Rules without Support Threshold: with and without Item Constraints, *TKDE*

Chiu, D.-Y., Wu, Y.-H. and Chen, A. L.P. (2004) An Efficient Algorithm for Mining Frequent Sequences by a New Strategy without Support Counting, *ICDE*

Dong, J., Perrizo, W., Ding, Q. and Zhou, J. (2000) The application of association rule mining to remotely sensed data, *Proceedings of the 2000 ACM symposium on Applied computing*

Fu, A. W.-C., Kwong, R. W.-W. and Tang, J. (2000). Mining n-most Interesting itemsets, *ISMIS*

Giannella, C., Han, J., Pei, J., Yan., X. and Yu, P. S. (2003) Mining Frequent Patterns in Data Streams at Multiple Time Granularities, *Next Generation Data Mining*

Gilburd,B., Schuster, A. and Wolff, R. (2004) A New Privacy Model and Association-Rule Mining Algorithm for Large-Scale Distributed Environments, *SIGKDD*

Grahne, G., Lakshmanan, L. and Wang, X. (2000) Efficient Mining of constrained correlated sets, *ICDE*

Han, J. and Fu, Y.(1995), Discovery of Multiple-Level Association Rules from Large Databases, *VLDB*

Han, J. and Kamber, M. (2000). Data Mining: Concepts and Techniques, *Morgan Kaufmann Publishers*.

Han, J., Pei, J. and Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation, *SIGMOD*

Han, J., Wang, J., Lu, Y. and Tzvetkov, P. (2002). Mining Top-K Frequent Closed Patterns without Minimum Support, *The 2002 IEEE International Conference on Data Mining (ICDM)*

Hedberg, S. (1995). The data gold rush, *BYTE, October*, 83-99

Hidber, C. (1999), Online association rule mining, *SIGMOD*

Kamber, M., Han, J. and Chiang, J. Y. (1997) Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes, *KDD*

Kleinberg, J., Papadimitriou, C. and Raghavan , P.(1998) A Microeconomic View of Data Mining, *Knowledge Discovery Journal*

Kuok, C.M., Fu, A.W.C. and Wong, M.H., (1998) Mining Fuzzy Association Rules in Databases *ACM SIGMOD Record, 27: 1*

Lee, W., Stolfo, S.J. and Mok, K.W. (1999) A data mining framework for building intrusion detection models, *IEEE Symposium on Security and Privacy*

Lent, B., Swami, A.N. and Widom, J. (1997) Clustering Association Rules, *ICDE*

Liu, B., Hsu, W. and Ma, Y. (1999) Mining association rules with multiple minimum supports, *SIGKDD*

Liu, B., Hsu, W. and Ma, Y. (1998) Integrating Classification and Association Rule Mining, KDD

Liu, G., Lu, H., Lou, W. and Yu, J. X. (2003) On Computing, Storing and Querying Frequent Patterns, *SIGKDD*

Liu, J., Pan, Y., Wang, K. and Han, J. (2002). Mining Frequent Item Sets by Opportunistic Projection, *SIGKDD*

Manku, G. S. and Motwani, R. (2002). Approximate Frequency Counts over Data Streams, *VLDB*

Ng, R., Lakshmanan, L.V.S., Han, J. and Pang, A. (1998) Exploratory Mining and Pruning Optimizations of constrained Association Rules, *SIGMOD*

Pei, J. and Han, J. (2000) Can we push more constraints into frequent pattern mining?, *KDD*

Pei, J., Han, J. and Mao, R. (2000). CLOSET: An efficient Algorithm for Mining Frequent Closed Itemsets, *DMKD*

Rizvi, S. J. and Haritsa, J. R. (2002) Maintaining Data Privacy in Association Rule Mining

Sarawagi, S., Thomas, S., Agrawal, R. (1998) Integrating association rule mining with relational database systems: alternatives and implications, SIGMOD

Sarda, N.L. and Srinivas, N.V. (1998) An Adaptive Algorithm for Incremental Mining of Association Rules, *Proceedings of the 9th International Workshop on Database and Expert Systems Applications*

Satou, K., Shibayama, G., Ono, T., Yamamura, Y., Furuichi, E., Kuhara, S. and Takagi, T. (1997) Finding association rules on heterogeneous genome data. *PSB*

Schuster, A., Wolff, R. and Trock, D (2003) A high-performance distributed algorithm for mining association rules, *ICDM*

Srikant, R. and Agrawal, R. (1995), Mining Generalized Association Rules, *VLDB*

Srikant, R. and Agrawal, R. (1996) Mining Quantitative Association Rules in Large Relational Tables, *SIGMOD*

Tao, F., Murtagh, F. and Farid, M. (2003) Weighted Association Rule Mining using weighted support and significance framework, *SIGKDD*

Vaidya, J. and Clifton, C. (2002) Privacy Preserving Association Rule Mining in Vertically Partitioned Data, *SIGKDD*

Wang, K. and Su, M.Y. (2002). Item Selection by "Hub-Authority" Profit Ranking, *SIGKDD*

Wong, R. C.-W., Fu, A. W.-C and Wang, K.(2003). MPIS: Maximal-Profit Item Selection with Cross-Selling Considerations, *The 2003 IEEE International Conference on Data Mining (ICDM)*

Wong, R.C.-W. and Fu, A.W.-C.(2004) ISM: Item Selection for Marketing with Cross-Selling Considerations, *PAKDD*

Yu, J. X., Chong, Z., Lu, H. and Zhou, A. (2004) False Positive or False Negative: Mining Frequent Itemsets from High Speed Transactional Data Streams, *VLDB*

Zaki, M. J. (1999), Parallel and Distributed Association Mining: A Survey, *IEEE Concurrency*

Zaki, M. J. and Hsiao, C. J. (2002) CHARM: An efficient Algorithm for Closed Itemset Mining, *SDM*

Zhang, H., Padmanabhan, B. and Tuzhilin, A. (2004) On the Discovery of Significant Statistical Quantitative Rules, *SIGKDD*

## TERMS AND THEIR DEFINITION

**Transaction:** a record containing the items bought by customers
**Apriori Algorithm:** an algorithm to find frequent itemsets and association rule. The process of finding frequent itemsets involves two steps – candidate itemset generation and pruning.
**FP-Growth:** an algorithm to find frequent itemsets by using a data structure called FP-tree
**Association Rule:** a kind of rules in the form $X \rightarrow I_j$, where $X$ is a set of some items and $I_j$ is a single item not in $X$.
**Support (Rule):** The support of a rule $X \rightarrow I_j$, where $X$ is a set of items and $I_j$ is a single item not in $X$, is the fraction of transactions containing all items in $X$ and item $I_j$.
**Confidence:** The confidence of a rule $X \rightarrow I_j$, where $X$ is a set of items and $I_j$ is a single item not in $X$, is the fraction of the transactions containing all items in set $X$ that also contain item $Ij$.
**Itemset:** a set of items
***k*-itemset:** itemset with k items
**Support (Itemset), or frequency:** The support of an itemset $X$ is the fraction of transactions containing all items in $X$.
**Frequent Itemset/Pattern:** the itemset with support greater than a certain threshold, called *minsupport*.
**Large Itemset:** same as Frequent Itemset. See Frequent Itemset.
**Infrequent Itemset:** the itemset with support smaller than a certain threshold, called *minsupport*.
**Small Itemset:** same as Infrequent Itemset. See Infrequent Itemset.
**Maximal Frequent Itemset/Pattern, or maxpattern:** frequent itemset which does not have a frequent item superset
**Frequent Closed Itemset/pattern:** an itemset $X$ if there exists no itemset $X'$ such that (1) $X \subset X'$ and (2) $\forall$ transactions $t$, $X \in t$ implies $X' \in t$.
**Loss Rule:** a kind of association rule used to model the loss of items in the problem MPIS
**Maximal-Profit Item Selection (MPIS):** the problem of item selection which selects a set of items in order to maximize the total profit with the consideration of cross-selling effect
**Boolean Association Rule:** association rule with only binary attributes
**Quantitative Association Rule:** association rule with quantitative attributes
**Single-level Association Rule:** a kind of association rule with one level of abstraction only
**Multi-level Association Rule:** a kind of association rule with different levels of abstractions