NEW METHOD TO MINING ASSOCIATION RULES USING MULTI-LAYER MATRIX QUADRANT

R.B. Fajriya Hakim
Statistics Department, Universitas Islam Indonesia
hakimf@fmipa.uii.ac.id

Abstract

Successful retail organizations utilizing any information they had for managing sales strategies. Most of information about consumer's retail organization had been stored in transactions database. Discovering knowledge from information stored in the transaction database has led several established methods implemented in many cases with their advantages and disadvantages. One of methodologies to uncover relationship among frequent items purchased in transaction database known as association rules. However, the research of association rules techniques to find knowledge from transaction database still provides a significant opportunity for new methods to participate. In this paper, we proposed a new method of mapping a frequent item set to a multi-layer matrix quadrant. This new method could show the metrics usually used to describe the association rules between items purchased same as any method used in association rules analysis.

Keywords: Association Rules, Matrix Quadrant, Support, Confidence, Lift Ratio

1. Introduction

Most large retail organizations store transactions data from day-to-day operations. Amounts of items purchased by customers usually recorded in the transaction databases. Those transactions database do not only show transaction date, unique identifier customer and the item purchased instead describing a lot of thing about customer behavior. Successful retail organizations view such databases as important part of information given by customer that could be utilized for managing sales strategies. Such supporting methodologies that can describe the behavior of consumers by using their items purchased are association rules analysis. This methodology is useful for discovering hidden interesting frequent items relationships in a transaction database. The relationships among frequent items purchased had been measured and presented in some known metrics of association rules analysis.

Mining association rules over transaction data was introduced in [1] and many established techniques has been recognized such as Apriori [2], FP-Tree [3] and [4], also hash-based association mining [5]. Some of their advantages and disadvantages have been uncovered by Han et al. [3], their application has been succeeded when it is implemented in Clustering on Sales Day Data [6], in very large clustered domains [7] and many other applications.

However, the research of association rules techniques to find knowledge from transaction database still provides a significant opportunity for new methods to participate. This paper is a pioneering work in a new technique that utilizes a multi-layer matrix quadrant to find association rules of transaction data.

2. Association Rules

Table 1 displays an example of simple market basket transactions data. There are five transactions performed by five different customers. Each row corresponds to a one transaction, which contains a unique identifier labeled TID (Transaction Identifier) followed by a set of items which is bought by a given customer. Curious retailers are interested in analyzing and find any valuable information from this transaction data to study purchasing behavior of their customers and utilizing for business marketing applications and promotions.

Table 1. Market Basket Transactions Data

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Coffee}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Coffee}

Such analysis and techniques to find valuable information from transaction data has been known as association analysis. These techniques are useful for discovering interesting relationships hidden in large data sets, but in this paper we only use a simple transaction data to describe our new technique. The uncovered relationships then could be described in the form of rules of association between sets of frequent items purchased. For example, some rules can be extracted from the transaction data set shown in Table 1:

$$\{\text{Bread}\} \rightarrow \{\text{Diapers}\},\ \{\text{Diapers}\} \rightarrow \{\text{Beer}\} \text{ or } \{\text{Bread}\} \rightarrow \{\text{Milk}\}$$

Those rules show there are strong relationships exist between some of items. The sale of bread and diapers, diapers and beer or bread and milk occurred because many customers who buy bread also buy diapers, who buy diapers also buy beer and who buy bread also buy milk. The rule of bread and milk maybe is a common behavior of customer's needs and does not more interesting than bread and diapers or diapers and beer. Those interesting rules as a knowledge discovered from transactions data could help retailers as a new opportunity to cross selling their products for their customers [8].

Association rules are a method that usually used in important applications such as market basket analysis to measure the associations between items which are purchased by consumers. In general, the objective of this method is to underline a set of items that typically occur together in a transaction. Data used in this method are taken from database of transactions, where each transaction (represented in a single row in the database) contains the list of items that purchased. Each customer may appear more than once in the database transaction, but in this market basket analysis a transaction means a single visit to the supermarket which the list of items purchased are recorded in the cashier. Rows typically contains a unique identifier of customers followed by a different number of items and detail of each items purchased. This is a remarkable difference numbers of items since we put it on data matrices. Based on this, the database can be converted in a binary data matrix, with transactions as rows and items as columns.

Transaction data usually can be represented in a binary format [8]. Simple market transaction data in table 1 are transformed into binary format as shown in Table 2, where itemset in each transaction have been collected and listed as a column title. The table shows that each row corresponds to a one transaction and each column corresponds to an item. The value is one if the item is present in a transaction and zero otherwise. This table could help in forming the matrix quadrant. Note that the column of this table should follow lexicographic order when it is used in our new technique.

Table 2. Binary Format Representation

TID	Bread	Milk	Diapers	Beer	Eggs	Coffee
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

The core problem of association rule technique in market basket analysis is to find, from the available transaction database, a subset of association rules that are interesting. We simply state

$$A \rightarrow B$$

to indicate an association rule of item A and item B. A is the antecedent, or body of the rule, and B is the consequent, or head of the rule. The rules are considered only to the sets of items (itemsets) that have transaction support count above minimum support. Support count refers to the number of transactions that contain a particular itemset. For a given rule, say $A \rightarrow B$, and $N_{A \rightarrow B}$ is the number of transactions in which the rule is satisfied. How strong an association rule between items could be measured in terms of its support and confidence [8]. The support for a rule $A \rightarrow B$ is obtained by dividing the number of transactions by the total number of transactions, N:

$$Support\ (A \rightarrow B) = \frac{N_{A \rightarrow B}}{N}$$

The support of a rule $A \rightarrow B$ is a relative frequency that indicates the proportion of A and B present together in data transactions, where A and B are disjoint itemsets. Bring this notation to probability rule, when a large sample is considered, the support will approximates the probability of occurrence:

$$Support\ (A \rightarrow B) = Probability\ (A \text{ and } B \text{ occur})$$

Support usually performed to filter out any rules which are obtained that are less frequent. Support is a simple and very useful measure of interestingness association between items.

The confidence of the rule $A \rightarrow B$ is obtained by dividing the number of transactions which item A and item B present in the transactions by the number of transactions which contain item A:

$$Confidence\ (A \rightarrow B) = \frac{N_{A \rightarrow B}}{N_A} = \frac{N_{A \rightarrow B} / N}{N_A / N} = \frac{Support\ (A \rightarrow B)}{Support\ (A)}$$

The confidence expresses how frequently items B appear in transactions that contain A.

The lift ratio for $A \rightarrow B$ is

$$Lift\ (A \rightarrow B) = \frac{Confidence\ (A \rightarrow B)}{Support\ (B)} = \frac{Support\ (A \rightarrow B)}{Support\ (A)\, Support\ (B)}$$

The Lift ratio is the confidence of the rule (A $\rightarrow$ B) divided by the support of B or we could say that the lift ratio is the support rule divided by each support item, assuming the support (item B) and the support (item A) are independent. Since a lift ratio is greater than

1.0, it suggests that we have a strong rule. Hence, this expresses that the larger the lift ratio, the greater is the strength of the association.

3. New Technique: Quadrant Matrix

The idea come up with the plotting the itemset transaction into matrix, Let **X** be a $2N$ x $2N$ matrix whose (i, j) entry is

$$X_{ij} = \begin{cases} \textit{transaction numbers}, & \textit{if } i \rightarrow j \\ \textit{empty}, & \textit{otherwise} \end{cases}$$

Note that we use $i \rightarrow j$ to mean that item i and item j are bought in a one transaction number, there may also be a different order of item j and item i, but this possibility is implied by the notation $i \rightarrow j$, since all items already ordered satisfying lexicographic order. The matrix is different with the common matrix, the matrix here is taken from Cartesian coordinates, where the Cartesian is partitioned into four equal quadrant which x and y axis contain the name of each item. Figure 1 shows the matrix quadrant.

				y-axis						
				...						
		Quadrant II		**Item3**			**Quadrant I**			
				Item2						
				Item1						
...	**-Item3**	**-Item2**	**-Item1**		**Item1**	**Item2**	**Item3**	**...**	**x-axis**	
				-Item1						
				-Item2						
		Quadrant III		**-Item3**			**Quadrant IV**			
				...						

Figure 1. Matrix Quadrant of Transaction Data

The entry of each cell is the transaction sequence numbers. Cells in the Quadrant I represent the purchased order items, items on the positive x-axis represent the first order and item on the positive y-axis represent the second order. The entry of cells in the main diagonal of $\boldsymbol{X}$ in the first quadrant represent to the item itself which was bought where the entry is also the transaction numbers. Cells in the Quadrant II represent continued purchase order items since the itemset transaction is more than two items i.e. three items and which positive y-axis represent second purchased item and negative x-axis represent the third item. The next quadrant follows that way. If the transaction contains more than five items and so on then the matrix could be expanded to multi-layer of matrix quadrant (Figure 2) and in each layer could be break apart for each quadrant into single matrix.

Note that the assumption of items purchased in database transaction will be sorted in lexicographic order. This assumption is important since there was no any information that the sequence of items should be preserved. Following (table 3) is the example of two transactions to explain our technique.

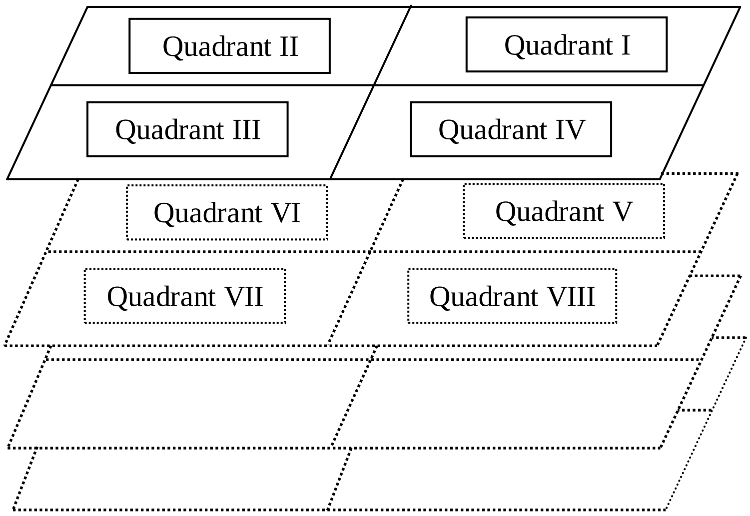

Figure 2. Multi-layer matrix quadrant of Transactional Data

Table 3. Example of Market Basket Transaction for Two Transaction Data

TID	Items
1	{A, B}
2	{A, D, C, E}

Data should be brought to lexicographic order,

TID	Items
1	{A, B}
2	{A, C, D, E}

Transform this lexicographic ordered items to the binary matrix (Table 4).

Table 4. Binary Format Representation of Two Transaction Data

TID	A	B	C	D	E
1	1	1	0	0	0
2	1	0	1	1	1

The technique start by inserting first transaction to the matrix quadrant (Figure 3)

		B	1	1
		A	1	
-B	**-A**		**A**	**B**
		-A		
		-B		

Figure 3. First transaction of example

Consider that the cell value in *A* and *A* are one (this value represents the first transaction), because *A* is on the list. The cell value in *B* and *B* also one, because *B* is on the list, then the value of *A* in the *x*-axis and *B* in the *y*-axis is one, because the first transaction is buying *A* and *B*. The second order of transactions cause the development of the matrix from 2 itemset to be 5 itemset.

					E	2		2	2	2
2,2					**D**	2		2	2	
2	2				**C**	2		2		
					B	1	1			
					A	1,2				
E	**D**	**C**	**B**	**A**		**A**	**B**	**C**	**D**	**E**
					A					
					B					
					C					
					D					
	2				**E**					

Figure 4. Second Sequence Transaction of example

First transaction only contains two different items i.e. *A* and *B*, while the second transaction contains three new items, that is, *C*, *D* and *E*. Hence, the total itemsets are *A*, *B*, *C*, *D* and *E*. The transactions data on the quadrant matrix will be look like at Figure 4.

Second sequence transaction data is {*A*, *C*, *D*, *E*} which means there is value in the cell of (*AA*, *AC*, *AD*, *AE*, *CC*, *CD*, *CE*, *DD*, *DE*, *EE*, *ACD*, *ACE*, *ADE*, *CDE*, *ACDE*) and partition on the first quadrant {*AA*, *AC*, *AD*, *AE*, *CC*, *CD*, *CE*, *DD*, *DE*, and *EE*} for the 2-itemset, on the second quadrant {*C*(-*D*), *C*(-*E*) and *D*(-*E*)} for the 3-itemset which represent {*ACD*, *ACE*, *ADE* and *CDE*} and on the third quadrant {(-*D*)(-*E*)} for the 4-itemset which represent {*ACDE*}. It is easy to explain the second sequence of transaction data on the quadrant matrix using arrow on the Figure 5.

					E	2		2	2	2
2, 2					D	2		2	2	
2	**2**				C	**2**		2		
					B	1	1			
					A	1,**2**				
E	**D**	C	B	A		**A**	B	C	D	E
					A					
					B					
					C					
					D					
	2				**E**					

Figure 5. Arrow Representation for The Transaction Data of {*A*, *C*, *D*, *E*}

Several notations from figure 5 could be explained here

a. Each item which is purchased, has entry cell in the main diagonal of first quadrant, e.g.: $X_{(item1, item1)} = X_{(A,A)} = \{1, 2\}$ means that the entry value on *x*-axis *A* and *y*-axis *A* or item *A* had been bought twice at the first and second transaction.

b. $X_{(C,\ (\ -D))}$ = {2} means that item *C* and item *D* on the second quadrant have been bought at the second transaction and has been preceded by item *A* purchased in the first quadrant.
c. $X_{((-\ D),\ (\ -E))}$ = {2} means that items *D* and *E* on the third quadrant have been bought at the second transaction and have been preceded by item *C* on the second quadrant and item *A* on the first quadrant.

Break apart the quadrant matrix to each matrix of each quadrant; we will get a matrix for first quadrant,

	A	***B***	***C***	***D***	***E***
A	12	1	2	2	2
B		1			
C			2	2	2
D				2	2
E					2

And a matrix for second quadrant,

	A	***B***	***C***	***D***	***E***
A					
B					
C				2	2
D					2
E					

And the last matrix for third quadrant,

	A	***B***	***C***	***D***	***E***
A					
B					
C					
D					2
E					

Matrix will be continued to the second layer of matrix quadrant if the frequent itemset transactions more than 5 items.

Back to the example on table 1, since there was no any information that the sequence of items in each transaction should be preserved, then the itemset data follow lexicographic orders. Only frequent purchased items are considered in association rules mining. Assume that those items listed are frequent items, the procedure to get the frequent items are just performing once scanning of transaction data that satisfying minimum support threshold.

TID	Items
1	{Bread, Milk}
2	{Beer, Bread, Diapers, Eggs}
3	{Beer, Coffee, Diapers, Milk}
4	{Beer, Bread, Diapers, Milk}
5	{Bread, Coffee, Diapers, Milk}

First transaction {Bread, Milk} is transfromed to the matrix quadrant (Figure 6)

		Milk	1	1
		Bread	1	
Milk	**Bread**		**Bread**	**Milk**
		Bread		
		Milk		

Figure 6. Matrix Quadrant for First Transaction

The element value of 1 in the cell of Bread and Bread, Milk and Milk, and Bread and Milk means that first transaction contains Bread, Milk and said that the first customer who buy Bread also buy Milk. In the quadrant I, the *x*-axis represents the first order and the *y*-axis represents the second order. The second sequence of transaction is {Beer, Bread, Diapers, Eggs}, hence the number of items in the *x*-axis and *y*-axis of matrix quadrant increased to Beer, Bread, Diapers, Eggs, Milk (lexicographic order). Figure 7 shows the transformation of second transaction data to the matrix quadrant.

					Milk		1			1
					Eggs	2	2	2	2	
	2				**Diapers**	2	2	2		
	2	2			**Bread**	2	12			
					Beer	2				
Milk	**Eggs**	**Diapers**	**Bread**	**Beer**		**Beer**	**Bread**	**Diapers**	**Eggs**	**Milk**
					Beer					
					Bread					
					Diapers					
		2			**Eggs**					
					Milk					

Figure 7. Second Sequence of Transaction Data

Figure 8 shows all transaction data from Table 1.

						Milk	34	145	35	345		1345
						Eggs	2	2		2	2	
334455	22					**Diapers**	234	245	35	2345		
35		35				**Coffee**	3	5	35			
4	2	24				**Bread**	24	1245				
						Beer	234					
Milk	**Eggs**	**Diapers**	**Coffee**	**Bread**	**Beer**		**Beer**	**Bread**	**Coffee**	**Diapers**	**Eggs**	**Milk**
						Beer						
						Bread						
						Coffee						
						Diapers						
		2				**Eggs**						
		345				**Milk**						

Figure 8. Matrix Quadrant for Transaction Data

Say, we would like to check the entry cell of {Bread, Diapers} = {2, 4, 5} in the Quadrant 1, then we continue to look at the second quadrant of the matrix, item Diapers in the positive *y*-axis was followed by {Diapers, Eggs} = {2, 2} and {Diapers, Milk} = {3,3,4,4,5,5}. Therefore we got the intersection

i. ({Bread, Diapers} = {2, 4, 5} $\cap$ {Diapers, Eggs} = {2, 2}) = {2}
ii. ({Bread, Diapers} = {2, 4, 5} $\cap$ {Diapers, Milk} = {3, 3, 4, 4, 5, 5}) = {4, 5}
The number of transaction that contains {Bread, Diapers, Eggs} is one which occurs in the second transaction. The support count for {Bread, Diapers, Eggs} equals to one because only one transaction which contains these three items. The number of transaction that contains {Bread, Diapers, Milk} is two which occur in the fourth and fifth transactions. The support count for these three items is two.

Consider the products Bread and Diapers and the support for the rule '*If Bread then Diapers*' which could be seen from quadrant first where *x*-axis is Bread and *y*-axis is Diapers. The number of transaction where Bread and Diapers purchased together is three which occur at second, fourth and fifth transaction

$$Support\ (Bread \rightarrow Diapers) = \frac{N_{Bread \rightarrow Diapers}}{N} = \frac{|\{Bread, Diapers\}|}{N} = \frac{|\{2,4,5\}|}{|\{1,2,3,4,5\}|} = \frac{3}{5} = 0.6$$

It is shown that the support of Bread and Diapers purchased together is 0.6. The confidence for the rule '*If Bread then Diapers*' can be explained by

$$Confidence\ (Bread \rightarrow Diapers) = \frac{N_{Bread \rightarrow Diapers}}{N_{Bread}} = \frac{N_{Bread \rightarrow Diapers}/N}{N_{Bread}/N} = \frac{\text{support}\ (Bread \rightarrow Diapers)}{\text{support}\ (Bread)} = \frac{0.6}{|\{1,2,4,5\}/\{1,2,3,4,5\}|} = \frac{0.6}{0.8} = 0.75$$

Therefore, it is shown that the proportion '*if Bread is bought then Diapers is also bought*' is 0.75.
And the lift ratio is

$$Lift\ (Bread \rightarrow Diapers) = \frac{confidence\ (Bread \rightarrow Diapers)}{\text{support}\ (Diapers)} = \frac{0.75}{|\{1,2,4,5\}/\{1,2,3,4,5\}|} = \frac{0.75}{0.8} = 0.9375 \text{ or}$$
$$= \frac{\text{support}\ (Bread \rightarrow Diapers)}{\text{support}(Bread)\ \text{support}(Diapers)} = \frac{0.6}{(0.8)\ (0.8)} = 0.9375$$

The lift ratio of 'if *Bread is bought then Diapers is also bought*' is 0.9375 it shows that the association between Bread and Diapers is not too strength in the transaction.
Another example, consider the rule

$$\{\text{Bread, Diapers}\} \rightarrow \{\text{Eggs}\}$$

{Bread, Diapers} $\rightarrow$ {Eggs} there are 3-itemset, the antecedent {Bread, Diapers} is on the Quadrant I and the consequence {Eggs} is on Quadrant II. Then we check at the Quadrant I of Matrix Quadrant, the cell of {Bread, Diapers} = {2, 4, 5} which is continued at the Quadrant II {Bread, Diapers} $\rightarrow$ {Eggs} = {2,2}, since the support count for {Bread, Diapers, Eggs}= is 1 then the support({Bread, Diapers} $\rightarrow$ {Eggs}) equal to = 0.2

(counted from {Bread, Diapers} = {2, 4, 5} and {Diapers, Eggs} = {2, 2} then {Bread, Diapers} $\cap$ {Diapers, Eggs} = {2, 4, 5} $\cap$ {2, 2} = {2} and the total number of transaction is 5, the rule's *Support* is 1/5 = 0.2). The rule's confidence is obtained by dividing the support count for {Bread, Diapers, Eggs} by the support count for {Bread, Diapers}. Since there are 3 transactions that contain Bread and Diapers, the *Confidence* for this rule is 0.2/(3/5) = 0.2/0.6 = 0.33. *Lift ratio* is *Confidence* {Bread, Diapers, Eggs} divided by *Support* (Eggs), since the *Support*(Eggs) = $N_{\{Eggs, Eggs\}}$ / N = {2}/{1,2,3,4,5} = 1/5 = 0.2 then the lift ratio of {Bread, Diapers} $\rightarrow$ {Eggs} is 0.33/0.2 = 1.665. Hence, it shows the strong association that 'if Bread and Diapers are bought then Eggs are bought'.

Using matrix quadrant to mining several metrics of association rules has shown that for some simple example this technique is working properly. This method differs from other methods of association rules because in this method the frequent itemset can be easily described in each quadrant, not only know the total rules formed but it can be known also how much of the association between items in each quadrant.

4. Conclusion

This paper explores a new technique in association rules, even though this paper only reveal a preliminary study but it could be extended in more general technique of mining association rules using multi-layer matrix quadrant. The simple example of transactional data has been shown with the result of association rules metrics such as support, confidence and lift ratio. Bring the transactions data to matrix will be easier to expand this technique to describe the formal definition and algorithm.

REFERENCES

1. Agrawal, R., Imielinski, T. And Swami, A., Mining Association Rules between Sets of Items in Large Databases. *In Proc. Of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., May 1993.
2. Agrawal, R. And Srikant, R., Fast Algorithms for Mining Association Rules, *Proceedings of the 20th VLDB Conference Santiago, Chile*, 1994
3. Han, J., Pei, J., And Yin, Y. Mining Frequent Patterns without Candidate Generation, *In Proc. ACM-SIGMOD Int Conf. Management of Data (SIGMOD'00)*, Dallas, TX, 2000: 1-12.
4. Han, J., Pei, J., And Yin, Y, Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree, *Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, 2004, 8: 53-87.
5. Park, J.S., Chen, M.S. And Yu, P.S., An Effective Hash-based Algorithm for Mining Association Rules. *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, 1995: 175-186.
6. Khattak, A.M., Kan, A.M., Lee, S., And Lee, Y.K., Analyzing Association Rule Mining and Clustering on Sales Day Data with XLMiner and Weka, *International Journal of Database Theory and Application*, March 2010, Vol.3. No.1.
7. Nanopoulos, A., Papadopoulos, A. N. And Manolopoulos, Y., Mining Association Rules in Very Large Clustered Domains, *Information Systems* 32 ,2007: 649-669.
8. Tan, Pang-Ning., Steinbach, Michael., Kumar, Vipin., *Introduction to Data Mining*, Addison-Wesley Longman, Boston, MA., USA., 2005.