BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **11**, No 2

Sofia • 2011

An Algebraic Representation of Frequent Market Baskets and Association Rules

J. Demetrovics[1]*, Hua Nam Son*[2]*, Akos Guban*[2]

[1] *MTA SZTAKI, 1111 Budapest, Lágymányosi u. 11*
[2] *Budapest Business School, 1149 Budapest, Buzogány u. 11-13*

Abstract: *This study proposes an algebraic approach for formal representation of Market Basket (MB) model. In a more generalized model by taking into consideration the quantity of items in transactions and by using tools of lattice theory we reconsider well-known problems and show an explicit representations of frequent MBs, basic frequent MBs and association rules. As straightforward consequences, the algorithms to find them are presented.*

Keywords: *market basket, frequent item, association rule, lattice.*

1. Introduction

Great efforts have been made to discover the informations hidden in the customer transactions. The study of customer Market Baskets (MB) and mining the association rules are important in various applications, for example, in decision making and strategy determination of retail economy [1]. In those studies the market baskets (transactions) are often considered as sets of items purchased by customers. Discovering of *large itemsets* and association rules attracts the interest of researchers. One can notice that in these studies the researchers are interested in the set of items (e.g. bread, milk, ...) purchased by customers in the super market, and did not care of the quantity of each item. However, it is interesting also if we know not only that 70% of customers buy bread and milk, but we know also 50% of customers buy 1 kg bread and 2 l milk, while 1% of customers buy 10 kg bread and 1 l milk. Similar example can be found for association rules. The meaning of quantitative analysis of transactions is evident.

In this study we introduce a quantitative analysis of transactions and association rules of transactions. The quantitative analysis may reveal informations hidden in the transactions. We are interested not only in the statement “90% of customers who buy bread and milk also purchase butter”, but in the statement “90% of customers who buy 1 kg bread and 2 l milk also purchase 0.5 kg butter”. By dealing with the quantity of items our setting is somehow different of those in previous studies (see [1]). That is why instead of *itemsets* (see [1]) we use *market baskets* or *transactions*. The main advantage of this approach is that all transactions can be examined as elements of a lattice with natural partial order. So the lattice-theoretic methods can be applied for transactions examination.

2. A generalized setting for Market Basket Model

For a finite set of items $P = \{p_1, p_2, ..., p_n\}$ we consider a MB as a tube $\alpha = (\alpha[1], \alpha[2], ..., \alpha[n])$, where $\alpha[i] \in \aleph$ is the quantity of p_i in the basket α. The set of all MBs is denoted by Ω.

For $\alpha, \beta \in \Omega$ where $\alpha = (\alpha[1], \alpha[2], ..., \alpha[n])$, $\beta = (\beta[1], \beta[2], ..., \beta[n])$ we write $\alpha \le \beta$ if for all $i = 1, 2, ..., n$ we have $\alpha[i] \le \beta[i]$. $\langle \Omega, \le \rangle$ is a lattice with the natural partial order $\le$. For a set $A \subseteq \Omega$ we denote

$$U(A) = \{\alpha \in \Omega \mid \forall \beta \in A : \beta \le \alpha\},$$

$$L(A) = \{\alpha \in \Omega \mid \forall \beta \in A : \alpha \le \beta\}.$$

We denote also

$$\sup(A) = \{\alpha \in U(A) | \nexists \beta \in U(A) \colon \beta < \alpha\},$$

$$\inf(A) = \{\alpha \in L(A) | \nexists \beta \in L(A) \colon \alpha < \beta\}.$$

One should remark that $\sup(A)$ and $\inf(A)$ are single elements of Ω, namely $\sup(A) = u \in \Omega$, where $u[i] = \max\{\alpha[i] | \alpha \in A\}$ and $\inf(A) = v \in \Omega$, where $v[i] = \min\{\alpha[i] | \alpha \in A\}$.

For a set $A \subseteq \Omega$ and $\alpha \in \Omega$ we denote by

$$\mathrm{supp}_A(\alpha) = \frac{|\{\beta \in A \mid \alpha \le \beta\}|}{|A|}$$

the support of α in A. In word, $\mathrm{supp}_A(\alpha)$ denotes the rate of all market baskets that exceeds the given threshold α (in the form of a sample market basket) to the whole A. The support of an market basket is a statistical index and naturally, the market baskets of more support are of more significance and attract the attention of the managers, as well as of the researchers.

One can notice that an item p_i (discused in other studies, see, for example, [1]) in our study should be identified with $U(\alpha_i)$, where $\alpha_i = (\alpha[1], \alpha[2], ..., \alpha[n])$, $\alpha[k] = 0$ if $k \ne i$ and $\alpha[i] = 1$. We should not confuse p_i with α_i.

For $\alpha, \beta \in \Omega$ where $\alpha = (\alpha[1], \alpha[2], ..., \alpha[n])$ and $\beta = (\beta[1], \beta[2], ..., \beta[n])$ we write $\gamma = \alpha \cup \beta$ if $\gamma[i] = \max\{\alpha[i], \beta[i]\}$ for all $i = 1, 2, ..., n$. We call

$\alpha \to \beta$ an *association rule* of β to α. By the *confidence* of $\alpha \to \beta$ in a set of MBs A we understand the rate

$$\operatorname{conf}_A(\alpha \to \beta) = \frac{\operatorname{supp}_A(\alpha \cup \beta)}{\operatorname{supp}_A(\alpha)}$$

As remarked in [1] the support of MBs is a kind of statistical index, while the confidence of association rules is a measure of their “strength”.

3. Frequent Market Baskets

For a set $A \subseteq \Omega$, $\alpha \in \Omega$ and $0 \le \varepsilon \le 1$ we say that α is ε-frequent MB, if $\operatorname{supp}_A(\alpha) \ge \varepsilon$. The set of all ε-frequent MBs is denoted by Φ_A^{ε}. We have the following

Apriori Principle. For a set $A \subseteq \Omega$, $\alpha, \beta \in \Omega$ and $0 \le \varepsilon \le 1$, if $\alpha \le \beta$ and β is ε-frequent then α is ε-frequent.

Example 1. Consider a set of items $P = \{a, b, c\}$ and a set of transactions $A = \{\alpha, \beta, \gamma, \delta\}$, where $\alpha = (2,1,0)$, $\beta = (1,1,1)$, $\gamma = (1,0,1)$, $\delta = (2,2,0)$. One can see that for $\sigma = (1,1,0)$, $\eta = (1,2,0)$ we have $\operatorname{supp}_A(\sigma) = \frac{3}{4}$ and $\operatorname{supp}_A(\eta) = \frac{1}{4}$. For the threshold $\varepsilon = \frac{1}{2}$ the ε-frequent MBs of A are:

$$\Phi_A^{\frac{1}{2}} = \{(2,1,0),(1,0,1),(1,1,0),(2,0,0),(0,0,1),(0,1,0),(1,0,0),(0,0,0)\}.$$

Let us denote

$$\Phi_{A,k} = \{\alpha \in \Omega \mid \exists \alpha_1, \alpha_2, ..., \alpha_k \in A : \alpha \le \{\alpha_1, \alpha_2, ..., \alpha_k\}\}.$$

One can remark that if $k \le l$ then $\Phi_{A,k} \supseteq \Phi_{A,l}$ and $\Phi_A^{\varepsilon} = \Phi_{A,k}$, where $k = \lceil \varepsilon \mid A \mid \rceil$ denotes the smallest integer that is greater or equal to $\varepsilon \mid A \mid$.

We have the following

Theorem 1. For a set of items $P = \{p_1, p_2, ..., p_n\}$, a set of MBs $A \subseteq \Omega$ and a threshold $0 \le \varepsilon \le 1$ an MB $\alpha \in \Omega$ is ε-frequent iff there exist $\alpha_1, \alpha_2, ..., \alpha_k \in A$ such that $\alpha \in L(\{\alpha_1, \alpha_2, ..., \alpha_k\})$, where $k = \lceil \varepsilon \mid A \mid \rceil$.

Proof: If there exist α_1, α_2, ..., $\alpha_k \in A$, $k = \lceil \varepsilon \mid A \mid \rceil$, such that $\alpha \in L(\{\alpha_1, \alpha_2, ..., \alpha_k\})$ then $\alpha \le \alpha_i$ for all $i = 1, 2, ..., k$, i.e.,

$$\operatorname{supp}_A(\alpha) = \frac{|\{\beta \in A \mid \alpha \le \beta\}|}{|A|} \ge \frac{k}{|A|} \ge \varepsilon .$$

Vice versa, if $\operatorname{supp}_A \ge \varepsilon$ then $|\{\beta \in A \mid \alpha \le \beta\}| \ge \varepsilon . |A|$, i.e. there exist $\alpha_1, \alpha_2, ..., \alpha_k \in A$, $k = \lceil \varepsilon \mid A \mid \rceil$, such that $\alpha \in L(\{\alpha_1, \alpha_2, ..., \alpha_k\})$. The proof is completed.

By the Theorem 1 we have the following

Algorithm 1 (Creating all *ε*-frequent MBs of a given set of transactions *A*).

```
Input. Set of items $P$, set of MBs $A \subseteq \Omega$ and a threshold $0 \le \varepsilon \le 1$.
Output. $\Phi_A^\varepsilon$.
    Step 1. $\Phi_A^\varepsilon := \varnothing$.
    Step 2. $k = \lceil \varepsilon \mid A \mid \rceil$.
        For all $B \subseteq A$, $\mid B \mid= k$
        $\Phi_A^\varepsilon := \Phi_A^\varepsilon \cup L(B)$
        EndFor;
    End
```

Let $\mid P \mid= n$, $k = \lceil \varepsilon \mid A \mid \rceil$, $m = max\{\alpha[i] \mid \alpha \in A, i = 1,2,...,n\}$. The algorithm requires $O\left(\binom{|A|}{k}.(m+1)^n\right)$ running time.

As a consequence of the previous theorem we have the following

Theorem 2 (Explicit representation of *large MBs*). For a set of items $P = \{p_1, p_2, ..., p_n\}$ a set of MBs $A \subseteq \Omega$ and a threshold $0 \le \varepsilon \le 1$ there exist $\alpha_1, \alpha_2, ..., \alpha_s \in \Omega,$ where $s = \binom{|A|}{\lceil \varepsilon |A| \rceil}$ such that

$$\Phi_A^\varepsilon = \bigcup_{i=1}^{s} L(\alpha_i).$$

Proof: Let $\alpha_1, \alpha_2, ..., \alpha_s$ be the set of all $\inf\{\beta_1, \beta_2, ..., \beta_k\}$ where $k = \lceil \varepsilon \mid A \mid \rceil$ and $\beta_i \in A$. By Theorem 1 we have

$$\alpha \in \Phi_A^\varepsilon \Leftrightarrow \alpha \le \inf(\{\beta_1, \beta_2, ..., \beta_k\})$$

for some $\{\beta_1, \beta_2, ..., \beta_k\} \subseteq A$, where $k = \lceil \varepsilon \mid A \mid \rceil$. This implies that $\Phi_A^\varepsilon = \bigcup_{i=1}^{s} L(\alpha_i)$. The proof is completed.

We should remark that $\alpha_i \le \alpha_j$ iff $L(\alpha_i) \subseteq L(\alpha_j)$. For a set of MBs A and a given threshold ε the set of MBs $\alpha_1,\ \alpha_2,\ ...,\ \alpha_s$ for which

(i) $\Phi_A^\varepsilon = \bigcup_{i=1}^{s} L(\alpha_i),$

(ii) $\forall i, j : 0 \le i, j \le s$ we have $\alpha_i \not\le \alpha_j$ and $\alpha_j \not\le \alpha_i$

is called by *basic ε-frequent set of MBs* of A. It is easy to verify that for a given A, ε the basic *ε*-frequent set of MBs of A is unique, which we denote by S_A^ε. Since the determination of Φ_A^ε (the set of all *ε*-frequent set of MBs in *A*) is important, it is interesting to determine its basic *ε*-frequent set of MBs S_A^ε. We have the following

Theorem 3. For a set of items *P*, a threshold $0 \le \varepsilon \le 1$ every set of MBs $A \subseteq \Omega$ has an unique basic ε-frequent set of MBs S_A^{ε}.

The simple proof is omitted. The following algorithm creates the unique basic ε-frequent set of MBs for a given set of MBs $A \subseteq \Omega$ and a given threshold ε:

Algorithm 2 (Creating the basic ε-frequent set of MBs S_A^{ε}).

Input. Set of items *P*, Set of MBs $A \subseteq \Omega$ and a thershold $0 \le \varepsilon \le 1$.

Output. S_A^{ε}.

```
Step 1. $S_A^{\varepsilon} := \varnothing$.
Step 2. $k = \lceil \varepsilon \,|\, A \,|\rceil$.
  For $B \subseteq A$, $|B| = k$
    For $\alpha \in S_A^{\varepsilon}$
      If $\alpha \le \inf(B)$ or $\inf(B) \le \alpha$ then
      $S_A^{\varepsilon} := S_A^{\varepsilon} \setminus \{\min(\alpha, \inf(B))\} \cup \{\max(\alpha, \inf(B))\}$.
      else
      $S_A^{\varepsilon} := S_A^{\varepsilon} \cup \{\inf(B))\}$.
      EndIf
    EndFor
  EndFor
End
```

For $|P| = n$, $k = \lceil \varepsilon \,|\, A \,|\rceil$, $m = \max\{\alpha[i] \,|\, i = 1, 2, ..., n; \alpha \in A\}$ one can see that $|S_A^{\varepsilon}| \le \binom{|A|}{k}$. Therefore the algorithm requires $O\left(\binom{|A|}{k} mn\right)$ running time. One can remark also than in the case of large amount of transactions A the basic ε-frequent set of MBs S_A^{ε} can be generated much more quickly than the set of all ε-frequent set of MBs Φ_A^{ε}.

Example 2. We continue the Example 1. For the set of transactions A Algorithm 2 generates the basic $\frac{1}{2}$-frequent set of MBs $S_A^{\frac{1}{2}} = \{\rho, \theta\}$, where $\rho = (2,1,0)$, $\theta = (1,0,1)$. It means that the family of $\frac{1}{2}$-frequent set of MBs of A is $\Phi_A^{\frac{1}{2}} = L(\rho) \cup L(\theta)$.

4. Association and confidence

In our generalized model of market baskets we can find all associations with given confidence. For a set of items *P*, a set of MBs $A \subseteq \Omega$ and a threshold $0 \le \varepsilon \le 1$ an association $\alpha \to \beta$ is *ε-confident* if $\mathrm{conf}_A(\alpha \to \beta) \geq \varepsilon$. The set of all ε-confident associations of A is denoted by C_A^{ε}. We have the following

Theorem 4. For a set of products P a set of MBs $A \subseteq \Omega$ and $0 \leq \varepsilon \leq 1$ an association $\alpha \to \beta$ is ε-confident iff $\dfrac{|U(\alpha \cup \beta) \cap A|}{|U(\alpha) \cap A|} \geq \varepsilon.$

Proof: Remark that $\mathrm{supp}_A(\alpha \cup \beta) = \dfrac{|U(\alpha \cup \beta) \cap A|}{|A|}$ and $\mathrm{supp}_A(\alpha) = \dfrac{|U(\alpha) \cap A|}{|A|}$. With these remarks the proof of the theorem is straightforward.

A natural question for cross marketing, store layout, ...(see, for example [1]) is to find all association rules with a given confidence. In our generalized model the following theorem shows in a sense an explicit representation of all association rules. More exactly, we show for a given MB α which set of MBs β may be associated to α with a given threshold of confidence.

For MBs ρ, σ where $\rho \leq \sigma$, let us denote

$$M(\rho,\ \sigma) = \{\eta \in \Omega \mid \rho \cup \eta \leq \sigma\}.$$

It should be remarked that $M(\rho, \sigma)$ can be represented explicitly. If $\rho = (\rho_1, \rho_2, \ldots, \rho_s)$, $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_s)$ then $\eta = (\eta_1, \eta_2, \ldots, \eta_s) \in M(\rho, \sigma)$ iff $\max(\rho_i, \eta_i)$ for all $i = 1, 2, ..., s$, i.e., $\eta_i = \sigma_i$ in the case $\rho_i \not\leq \sigma_i$ and $\eta_i \leq \sigma_i$ in the case $\rho_i = \sigma_i$.

Theorem 5 (Explicit representation of association rules). For a set of items $P = \{p_1, p_2, ..., p_n\}$, a set of MBs $A \subseteq \Omega$, an MB $\alpha \in \Omega$ and a threshold $0 \leq \varepsilon \leq 1$ there exist $\alpha_1, \alpha_2, ..., \alpha_k \in \Omega$ such that $\forall \beta \in \Omega : \alpha \to \beta$ is ε-confident association rule iff $\beta \in \bigcup_{i=1}^{k} M(\alpha, \alpha_i)$.

Proof: Put $s = \lceil \varepsilon |U(\alpha) \cap A| \rceil$ by Theorem 4 we have that $\alpha \to \beta$ is ε-confident association rule iff $|U(\alpha \cup \beta) \cap A| \geq s$. Let α_i denotes $\inf(B)$, where $B \subseteq A$, $|B| \geq s$. One can verify that $|U(\alpha \cup \beta) \cap A| \geq s$ iff $\beta \in M(\alpha, \alpha_i)$. The proof is completed.

Theorem 5 in a sense gives an explicit presentation for association rules. As a straightforward consequence, we have an algorithm to find all ε-confident association rules for given left side.

Algorithm 3 (Creating all ε-confident association rules $\alpha \to \beta$ for given α).

Input. A set of items *P*, a set of MBs $A \subseteq \Omega$, a thershold $0 \leq \varepsilon \leq 1$ and an MB α

Output. $\bigcup_{i=1}^{k} M(\alpha, \alpha_i)$.

Step 1. $B := U(\alpha) \cap A = \{\gamma \in A \mid \alpha \leq \gamma\}$.

```
Step 2. $s := \lceil \varepsilon | B | \rceil$.
  $k := | \{C \subseteq B \| C | \geq s\} |$
  For $C \subseteq B$, $| C | \geq s$, calculate $\alpha_i = \inf(C)$, $i = 1, 2, ..., k$.
  EndFor
Step 3.
  For $i = 1, 2, ..., k$ calculate $M(\alpha, \alpha_i)$
  EndFor
Step 4.
  Output $\bigcup_{i=1}^{k} M(\alpha, \alpha_i)$.
End
```

Example 3. We continue the Example 1. For the set of MBs A (see Example 1), the MB $\sigma = (1, 1, 0)$ and threshold $\varepsilon = \frac{1}{2}$ we should find all MB η such that $\sigma \to \eta$ is ε-confident association rule. We can see $U(\sigma) \cap A = \{(2, 1, 0), (1, 1, 1), (2, 2, 0)\}$ and $s := \lceil \varepsilon | U(\alpha) \cap A | \rceil = 2$. By Step 2 in Algorithm 3 we have $k = 4$ and $\alpha_1 = (1, 1, 0)$, $\alpha_2 = (2, 1, 0)$. The set of all MBs η such that $\sigma \to \eta$ is $\frac{1}{2}$- confident association rule is

$$M(\sigma, \alpha_1) \cup M(\sigma, \alpha_2) = \{(1, 1, 0), (1, 0, 0), (0, 1, 0), (0, 0, 0), (2, 1, 0), (2, 0, 0)\}.$$

As a result we see that besides the trivial association rules of the form $\sigma \to \sigma'$, where $\sigma' \leq \sigma$ we got non-trivial association rules $\sigma \to (2, 1, 0)$ and $\sigma \to (2, 0, 0)$. In words, among those customers A the ratio of customers who buy a and b also buy two a and one b items, as well the ratio of those who buy a and b also buy two a items, are more than 50%.

5. Conclusion

In this study we have proposed an algebraic approach to consider the MB model. The well-known problems are analysed in new, more generalized setting. An explicit representation of frequent set of MBs, as well of association rules are presented. We define the set of basic frequent MBs which determines the set of frequent MBs and can be created in shorter time. We described algorithms that produces the set of frequent MBs and the set of basic frequent MBs. We described also an algorithm that produces the set of association rules for a given left side. The algebraic approach we propose here brings about a clearer representation of well-known results and appears to be a good tool for future study in market basket model.

References

1 Agrawal, R., R. Srikant. Fast Algorithms for Mining Association Rules. VLDB, 1994, 487-499.

2. Brũggermann, T., P. Hedstrõm, M. Josefsson, Data Mining and Data Based Direct Marketing Activities, Book on Demand GmbH. Norderstedt, Germany, 2004.

3. Han, J., M. Kamber. Data Mining: Concepts and Techniques. Second Edition. Morgan Kaufmann Publ., 2006.

4. Mannila, H., H. Toivonen. Discovering Generalized Episodes Using Minimal Occurrences. – In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD' 96). August 1996, AAAI Press, 146-151.

5. Toivonen, H. Sampling Large Databases for Association Rules. Morgan Kaufmann Publ., 1996, 134-145.

6. Pasquier, N., Y. Bastide, R. Taouil, L. Lakhal. Discovering Frequent Closed Itemsets for Association Rules – ICDT, 1999, 398-416.

7. Hsu, Ping-Yu, Yen-Liang Chen, Chun-Ching Ling. Algorithms for Mining Association Rules in Bag Databases. – Information Sciences, Vol. **166**, 2004, Issues 1-4, 31-47.