Annales Univ. Sci. Budapest., Sect. Comp. 42 (2014) 135–156

AN ALGEBRAIC APPROACH TO THE STUDY OF MARKET BASKETS AND THEIR CLASSIFICATION

J. Demetrovics (Budapest, Hungary)

Hua Nam Son (Budapest, Hungary)

A. Guban (Budapest, Hungary)

Dedicated to András Benczúr on the occasion of his 70th birthday

Communicated by Péter Racskó

(Received June 1, 2014; accepted July 1, 2014)

Abstract. The paper focuses on the algebraic representation of market basket model. It is shown in this paper that the methods offered by the new approach are effective in analyzing the problems concerning the customer's market baskets. The results of the previous studies in discovering the frequent market baskets and the association rules between market baskets, as well as the definition of the constraints of market baskets are summarized. By using these methods the logical structure of the sets of market baskets is analysed and the complexity of the market baskets is determined. In this formalism this paper shows also that the algebraic model of market baskets is quite suitable for solving the problems concerning the market basket's classification. The operations on classifications are discussed. A new concept of neighborhood between market baskets is introduced and their properties are studied in this paper. It should be remarked that in this algebraic model the customers and the transactions can be identified by their market baskets. This implies that the results that hold for market baskets hold also for customers and transactions. The logical and algebraic methods that have been used to study the frequent market baskets, the association rules between market baskets and the constraints of market baskets here appear to be efficient tools in the study of the customer's classification.

Key words and phrases: frequent itemset, association rule, algorithm, customer classification.

1. Introduction

Discovering the hidden informations in the sets of market baskets and in the sets of customer's transactions is always interesting problem that has attracted the attention of researchers (see, for example, [1, 3, 7, 8]). The studies of customer market baskets (MBs) and mining the frequent itemsets, as well as the association rules are important in different applications, for example, in decision making and strategy determination of retail economy ([1]). As noticed in previous researches ([4]) most of the studies concerning the market baskets dealt with only the set of items purchased by customers or involved in the transactions. The quantity of the items in transactions were not considered and therefore its important role in the analysis of transactions were ignored. Here the market baskets and the association rule between market baskets are studied in more details: instead of discovering the association rule between wheat flour and egg, or between bread and milk, the association rule between 1 kg wheat flour and 10 pieces of egg, or the association rule between 1 kg bread and 2 liter of milk are studied. It would be remarked that among those customers who buy wheat flour and eggs most of them buy 1 kg wheat flour and 10 pieces of egg, while the least of them buy 10 kg wheat flour and 1 piece of egg. Evidently, the quantitative analysis is necessary.

In Section 2 we recall the algebraic formalism for analysis of market baskets which was established firstly in [4]. In Section 3 the results related to frequent MBs and association rules are resumed. The structure of frequent MBs and association rules are shown. The concept of the constraint of MBs is introduced in Section 4. It is shown that every set of MBs can be characterized by some logical formula that is called by the constraint of MBs. The dependencies between MBs as special form of constraints are induced in a natural way by the implications between logical formulas. Based on the results concerning the constraints of MBs in Section 5 we introduce the concept of complexity of the sets of MBs. The concepts and problems in the classification of MBs are proposed and studied in Section 6. Some aspects and open problems are discussed in the conclusion in Section 7.

2. Market basket model

In this section the concepts and results previously established in the formalism in [4] are recalled. Let $P = \{p_1, p_2, ..., p_n\}$ be a finite set of items. A *market*

basket (MB) is a tuple $\alpha = (\alpha[1], \alpha[2], ..., \alpha[n])$, where $\alpha[i] \in \mathbb{N}$ is the quantity of the item p_i in the basket. The set of all MBs is denoted by Ω. We can remark:

1. By the condition $\alpha[i] \in \mathbb{N}$ in the definition of market baskets we can see that in the previous studies, as well as in this study only the market baskets with integer components are considered. A more generalized model where the market baskets with components being real numbers, $\alpha[i] \in \mathbb{R}$, may be interesting topics of other study.

2. In the case of market baskets with components being integers the market baskets can be considered as vectors with integer components. This enables us to study different structures on these market baskets.

3. A customer or a transaction in fact can be identified as a market basket. Thus in the followings the concepts and results concerning market baskets in this sense hold also for the customers, transactions. The problems concerning the customers, for examples, the determination of frequent customers, the association rules between customers, etc., are interesting problems in practice.

By the previous remarks let us consider a structure on the set of MBs Ω. For $\alpha, \beta \in \Omega$ where $\alpha = (\alpha[1], \alpha[2], ..., \alpha[n])$, $\beta = (\beta[1], \beta[2], ..., \beta[n])$ we write $\alpha \leq \beta$ if for all $i = 1, 2, ..., n$ we have $\alpha[i] \leq \beta[i]$. $\langle \Omega, \leq \rangle$ is a lattice with the natural partial order $\leq$. For a set $A \subseteq \Omega$ we denote by $U(A)$, $L(A)$ the set of all upper, or lower bounds of A, respectively: $U(A) = \{\alpha \in \Omega | \forall \beta \in A : \beta \leq \alpha\}$ and $L(A) = \{\alpha \in \Omega | \forall \beta \in A : \alpha \leq \beta\}$.

We denote also by $sup(A)$ and $inf(A)$, respectively, the smallest, and the largest MB in $U(A)$ and $L(A)$.

The *support* of an MB $\alpha \in \Omega$ in a set of MBs $A \subseteq \Omega$ is defined as the proportion

$$supp_A(\alpha) = \frac{|\{\beta \in A | \alpha \leq \beta\}|}{|A|},$$

that is in fact the rate of all MBs in A exceeding the given sample MB α to the whole A. In other words, $supp_A(\alpha)$ denotes the proportion of those customers who "support" α to the whole group of customers A. Here one can see the double meaning of MBs: MBs on the one hand are viewed as itemsets, on the other hand they are considered as customers. Naturally, discovering of the highly supported MBs is an important problem in various areas of economy.

3. Frequent itemsets and association rules

For a set of MBs $A \subseteq \Omega$, an MB $\alpha \in \Omega$ and for a threshold $0 \leq \varepsilon \leq 1$ the ε-frequent MBs are those MBs whose support exceeds ε, i.e. if $supp_A(\alpha) \geq \varepsilon$. The set of all ε-frequent MBs is denoted by Φ_A^ε.

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 $\alpha \longrightarrow \beta$ an *association rule*. By the *confidence* of $\alpha \longrightarrow \beta$ in a set of MBs A we mean the proportion

$$conf_A(\alpha \longrightarrow \beta) = \frac{supp_A(\alpha \cup \beta)}{supp_A(\alpha)}.$$

The following example was considered in [4]:

Example 3.1. *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 $supp_A(\sigma) = \frac{3}{4}$ and $supp_A(\eta) = \frac{1}{4}$. 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

$$\Phi_{A,k} = \{\alpha \in \Omega | \exists \alpha_1, \alpha_2, ..., \alpha_k \in A, \alpha_i \neq \alpha_j \, (i \neq j) : \alpha \leq \{\alpha_1, \alpha_2, ..., \alpha_k\}\}.$$

One can remark that if $k \leq l$, then $\Phi_{A,k} \supseteq \Phi_{A,l}$ and $\Phi_A^\varepsilon = \Phi_{A,k}$ where $k = \lceil \varepsilon |A| \rceil$ denotes the smallest integer that is greater or equal to $\varepsilon |A|$. The following Theorem 3.2, 3.3 were proved in [4]:

Theorem 3.2. *For a set of items $P = \{p_1, p_2, ..., p_n\}$, a set of MBs $A \subseteq \Omega$ and a threshold $0 \leq \varepsilon \leq 1$ 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 $k = \lceil \varepsilon |A| \rceil$.*

By Theorem 3.2 in [4] an algorithm was proposed that creates all ε-frequent MBs for a given set of MBs A in $O\left(\binom{|A|}{k} . (m+1)^n\right)$ running time.

Algorithm 3.1: (Creating all ε-frequent MBs of a given set MBs A)

Input: Set of items P, Set of MBs $A \subseteq \Omega$ and a threshold $0 \leq \varepsilon \leq 1$.

Output: Φ_A^ε.

Theorem 3.3. *(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* $0 \leq \varepsilon \leq 1$ *there exist* $\alpha_1, \alpha_2, ..., \alpha_s \in \Omega$ *where* $s = \binom{|A|}{\lceil \varepsilon |A| \rceil}$ *such that*

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

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

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

(ii) $\forall i, j : 0 \leq i, j \leq s$ we have $\alpha_i \nleqq \alpha_j$ and $\alpha_j \nleqq \alpha_i$.

For a given A, ε the basic ε- frequent set of MBs of A is unique, which we denote by S_A^ε. We have

Theorem 3.4. *For a set of items* P*, a threshold* $0 \leq \varepsilon \leq 1$ *every set of MBs* $A \subseteq \Omega$ *has a unique basic ε- frequent set of MBs* S_A^ε.

An algorithm that creates the basic ε- frequent set of MBs in $O\left(\binom{|A|}{k}.m.n\right)$ running time for a given set of MBs $A \subseteq \Omega$ and a given threshold ε is proposed in [4]:

Algorithm 3.2: (Creating the basic ε- frequent set of MBs S_A^ε)

Input: Set of items P, Set of MBs $A \subseteq \Omega$ and a threshold $0 \leq \varepsilon \leq 1$.

Output: S_A^ε.

One can remark that in the case of large amount of MBs 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 3.5. *We continue the Example 3.1. For the set of transactions A the Algorithm 3.2 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 $\Phi_A^{\frac{1}{2}} = L(\rho) \cup L(\theta)$.*

As shown in [4] we can find all associations with given confidence. For a set of items P, a set of MBs $A \subseteq \Omega$ and a threshold $0 \leq \varepsilon \leq 1$ an association $\alpha \longrightarrow \beta$ is *ε-confident* if $conf_A(\alpha \longrightarrow \beta) \geq \varepsilon$. The set of all ε-confident associations of A is denoted by C_A^ε. We have

Theorem 3.6. *For a set of items P, a set of MBs $A \subseteq \Omega$ and $0 \leq \varepsilon \leq 1$ an association $\alpha \longrightarrow \beta$ is ε-confident iff* $\dfrac{|U(\alpha \cup \beta) \cap A|}{|U(\alpha) \cap A|} \geq \varepsilon$.

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

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

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 $\eta = (\eta_1, \eta_2, \ldots, \eta_s) \in M(\rho, \sigma)$ if and only if $max(\rho_i, \eta_i) = \sigma_i$ for all $i = 1, 2, \ldots, s$, i.e. $\eta_i = \sigma_i$ in the case $\rho_i \lneq \sigma_i$ and $\eta_i \leq \sigma_i$ in the case $\rho_i = \sigma_i$.

Theorem 3.7. *(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 $\forall \beta \in \Omega : \alpha \longrightarrow \beta$ is an ε-confident association rule if and only if $\beta \in \bigcup_{i=1}^{k} M(\alpha, \alpha_i)$.*

As we have shown in [4] Theorem 3.7 in a sense gives an explicit presentation for association rules and by the following algorithm one can find all ε-confident association rules for given left side.

Algorithm 3.3: (Creating all ε- confident association rules $\alpha \longrightarrow \beta$ for given α)

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

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

Example 3.8. *We continue the Example 3.1. For the set of MBs A the MB $\sigma = (1, 1, 0)$ and threshold $\varepsilon = \frac{1}{2}$ we should find all MB η such that $\sigma \longrightarrow \eta$ 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 3.3 we have $k = 4$ and $\alpha_1 = (1, 1, 0)$, $\alpha_2 = (2, 1, 0)$. The set of all MBs η such that $\sigma \longrightarrow \eta$ 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 result we see that besides the trivial association rules of the form $\sigma \longrightarrow \sigma^{'}$, where $\sigma^{'} \leq \sigma$ we got non-trivial association rules $\sigma \longrightarrow (2,1,0)$ and $\sigma \longrightarrow (2,0,0)$. In words, among those customers A the ratio of customers who buy a and b also buy 2 a and 1 b items, as well the ratio of those who buy a and b also buy 2 a items, are more than 50 percent.

4. Constraints of market baskets

In this section we consider the constraints of MBs. As introduced previously in [4] by *constraints* of MBs we understand the logical formula that represent these MBs. For example, the constraint $(\neg\alpha)$ where α means the meat certainly holds with high support for the vegetarian customer's groups. In the same way, the constraint $(\alpha \wedge \beta) \longrightarrow \gamma$ seemingly gains high support from the householder customers, if α, β and γ means milk, egg and wheat flour respectively. By the dependency between MBs we can understand the logical implication of the form $\alpha \longrightarrow \beta$ that in fact are special constraints.

Let us construct the logical constraints of MBs. For a set of items $P = \{p_1, p_2, ..., p_n\}$ let Ω be the set of all MBs over P. We define the *logical constraints of MBs* (for short, constraint) as follows:

(1) All $\alpha \in \Omega$ are constraints. In this case $\pi(\alpha) = U(\alpha) = \{\beta \in \Omega | \alpha \leq \beta\} \subseteq \Omega$.

(2) If α is a constraint, then $(\neg\alpha)$ is a constraint and $\pi(\neg\alpha) = (\pi(\alpha))^c$ where by A^c we denote $\Omega \setminus A$ for $A \subseteq \Omega$.

(3) If α, β are constraints, then

$(\alpha \vee \beta)$ is a constraint and $\pi(\alpha \vee \beta) = \pi(\alpha) \cup \pi(\beta)$,

$(\alpha \wedge \beta)$ is a constraint and $\pi(\alpha \wedge \beta) = \pi(\alpha) \cap \pi(\beta)$.

(4) All constraints are constructed as in 1., 2. and 3.

As usual, the parentheses are omitted where it causes no confusion. We call $\pi(\alpha)$ the *set of supporting market baskets* of α. Two constraints α, β are *equivalent*, noted by $\alpha \equiv \beta$, if $\pi(\alpha) = \pi(\beta)$. A constraint is *tautology* if $\pi(\alpha) = \Omega$. The set of all constraints is denoted by $C(\Omega)$.

The following properties of propositions in propositional calculus hold also for the constraints:

(1) If α, β, $\gamma \in C(\Omega)$ are constraints, then

$\alpha \vee \beta \equiv \beta \vee \alpha$, $\quad \alpha \wedge \beta \equiv \beta \wedge \alpha$,

$\alpha \vee (\beta \vee \gamma) \equiv (\alpha \vee \beta) \vee \gamma$, $\quad \alpha \wedge (\beta \wedge \gamma) \equiv (\alpha \wedge \beta) \wedge \gamma$.

(2) If $\alpha \in C(\Omega)$ is a constraint, then $\neg(\neg\alpha) \equiv \alpha$.

(3) If α, $\beta \in C(\Omega)$ are constraints, then

$\neg(\alpha \wedge \beta) \equiv \neg\alpha \vee \neg\beta$ and

$\neg(\alpha \vee \beta) \equiv \neg\alpha \wedge \neg\beta$.

(4) For α, $\beta \in C(\Omega)$ the notation $\alpha \to \beta$ is used also for $\neg\alpha \vee \beta$.

The above identities are always true. We call these identities the *logical identities*. It is easy to see that for a given A in the same way we can define $\pi_A(\alpha) = \pi(\alpha) \cap A$, which we call the *relative set of supporting MBs* of α. Similarily we say that two constraints α, β are *relatively equivalent* (in A), noted by $\alpha \equiv_A \beta$, if $\pi_A(\alpha) = \pi_A(\beta)$. It is easy to verify the following

Theorem 4.1. (1) *For any finite set of MBs $A \subseteq \Omega$ there is a constraint $\alpha_A^* \in C(\Omega)$ such that $\pi(\alpha_A^*) = A$.*

(2) *For all $\beta, \gamma \in C(\Omega)$, $\beta \equiv_A \gamma$ if and only if $\beta \wedge \alpha_A^* \equiv \gamma \wedge \alpha_A^*$.*

Proof.

1) For any finite set of MBs $A \subseteq \Omega$ we find the constraint $\alpha_A^* \in C(\Omega)$ such that $\pi(\alpha_A^*) = A$. If $P = \{p_1, p_2, ..., p_n\}$, $\rho = (\rho[1], \rho[2], ..., \rho[n]) \in \Omega$ then let

$$\rho_i^+ = (\rho[1], \rho[2], ..., \rho[i] + 1, ..., \rho[n]).$$

One can see that

$$\{\rho\} = \pi(\rho) \setminus \bigcup_{i=1}^{n} \pi(\rho_i^+) = \pi(\rho \wedge \bigwedge_{i=1}^{n} \neg(\rho_i^+)).$$

Let

$$\alpha_A^* = \bigvee_{\rho \in A} [\rho \wedge \bigwedge_{i=1}^{n} \neg(\rho_i^+)].$$

We have $A = \pi(\alpha_A^*)$.

2) The assertion is proved easily by using the definitions. We have $\beta \equiv_A \gamma$ $\iff \pi_A(\beta) = \pi_A(\gamma) \iff \pi(\beta) \cap A = \pi(\gamma) \cap A \iff \beta \wedge \alpha_A^* \equiv \gamma \wedge \alpha_A^*$. ■.

One can remark that there are two trivial cases: The first one is the case, when α_A^* is tautology. In this case $\equiv_A$ coincides with $\equiv$, which does not hold in general. We call a set of customers (transactions) *complete* if α_A is tautology. The second case is when α_A^* is tautologically false. For $\beta \in C(\Omega)$ we denote $\beta_A = \beta \wedge \alpha_A^*$.

Example 4.2. *We continue the Example 3.1. Let* $P = \{a, b, c\}$ *and a set of transactions* $A = \{\alpha, \beta, \gamma, \delta\}$*, where* $\alpha = (2, 1, 0)$*,* $\beta = (1, 1, 1)$*,* $\gamma = (1, 0, 1)$*,* $\delta = (2, 2, 0)$*. If* a *= "Flour",* b *= "Egg",* c *= "Milk", which can be identified by* $a = (1, 0, 0)$*,* $b = (0, 1, 0)$*, and* $c = (0, 0, 1)$*, respectively, then*

$$\begin{array}{ll} \pi(a) = U((1,0,0)) = \{(x,y,z) | x \geq 1\}, & \pi_A(a) = \{\alpha, \beta, \gamma, \delta\}, \\ \pi(b) = U((0,1,0)) = \{(x,y,z) | y \geq 1\}, & \pi_A(b) = \{\alpha, \beta, \delta\}, \\ \pi(c) = U((0,0,1)) = \{(x,y,z) | z \geq 1\}, & \pi_A(c) = \{\beta, \gamma\}. \end{array}$$

In this case the constraint $a \wedge b \to c$ that may be interpreted as $Flour \wedge Egg \to Milk$, characterises those customers, who if buy Flour and Egg then must buy Milk. It is easy to see that the set of supporting MBs of this constraint is $\pi(a \wedge b \to c) = \{(x, y, z) | x = 0 \text{ or } y = 0 \text{ or } z \geq 1\}$. One also can see that in this case $\pi_A(a \wedge b \to c) = \pi(a \wedge b \to c) \cap A = \{\beta, \gamma\}$, i.e. $(a \wedge b \to c) \equiv_A c$.

It is easy to see that the properties of propositions in propositional calculus hold also for the constraints in the given set of customers, but the converse is not always true. Although one can verify the followings for α, $\beta \in C(\Omega)$ and an arbitrary set of customers A:

(1) $(\alpha \vee \beta)_A \equiv_A \beta_A \vee \alpha_A$,

(2) $(\alpha \wedge \beta)_A \equiv_A \beta_A \wedge \alpha_A$,

(3) $(\neg\alpha)_A \equiv_A \neg(\alpha_A)$.

One should distinguish $\equiv_A$ and $\equiv$.

5. The complexity of market baskets

In this section we propose a criteria for the complexity of customer sets. The practical aspect of this attempt is clear: every shop manager wants to know how complex his customer set is or how his customer set should be classified into groups. One can remark that the set of customers that contains only one

customer is simple. An other simple customer set is the case when the transactions of the customers in the set (that may be a large mass) are "similar". The concept of complexity of customer sets may be understood as followings.

Let $P = \{p_1, p_2, ..., p_n\}$ be a finite set of items and Ω be the set of MBs over P. We recall that $U(\alpha) = \{\beta \in \Omega | \alpha \leq \beta\}$ for $\alpha \in \Omega$. We call a set $B \subseteq \Omega$ a *block of customers* if there are $\alpha_1, \alpha_2, \ldots, \alpha_m \in \Omega$; $\beta_1, \beta_2, \ldots, \beta_n \in \Omega$ such that

$$B = \bigcap_{k=1}^{m} U(\alpha_k) \setminus \bigcup_{k=1}^{n} U(\beta_k).$$

The block is denoted by $[\alpha_1, \alpha_2, \ldots, \alpha_m | \beta_1, \beta_2, \ldots, \beta_n]$. We have the following simple theorem.

Theorem 5.1. *Let $P = \{p_1, p_2, ..., p_n\}$ be a a finite set of items and Ω be the set of all MBs over P.*

(1) *Every $\gamma \in \Omega$ is a block, i.e. there are $\alpha_1, \alpha_2, \ldots, \alpha_m \in \Omega$; $\beta_1, \beta_2, \ldots, \beta_n \in \Omega$ such that $\{\gamma\} = [\alpha_1, \alpha_2, \ldots, \alpha_m | \beta_1, \beta_2, \ldots, \beta_n]$.*

(2) *Every $A \subseteq \Omega$ is union of some blocks, i.e. there are $0 \leq k$, $\alpha_1^k, \alpha_2^k, \ldots, \alpha_{m_k}^k \in \Omega$, $\beta_1^k, \beta_2^k, \ldots, \beta_{n_k}^k \in \Omega$ such that*

$$A = \bigcup_{i=1}^{k} [\alpha_1^i, \alpha_2^i, \ldots, \alpha_{m_i}^i | \beta_1^i, \beta_2^i, \ldots, \beta_{n_i}^i].$$

Let us denote

$$c(A) = \min \left\{ k \mid \exists B_k \text{ blocks, such that } A = \bigcup_{i=1}^{k} B_i \right\}.$$

$c(A)$ can be considered as a kind of the *complexity* of A. If $A = \bigcup_{i=1}^{k} B_i$ where $k = c(A)$ then we say that $A = \bigcup_{i=1}^{k} B_i$ is a *minimal representation* of A by blocks. We should notice that a set $A \subseteq \Omega$ may have different minimal representations, even if we does not take in account of the permutation of blocks. Let us consider an example.

Example 5.2. *Following the Example 4.2 let $\alpha = (2, 1, 0)$, $\beta = (1, 1, 1)$, $\gamma = (1, 0, 1)$ and let $\theta = (1, 1, 2)$, $\lambda = (1, 0, 2)$. One can verify*

$$\{\gamma\} = U(\gamma) \setminus \{U((2, 0, 1)) \cup U(\beta) \cup U(\lambda)\}$$

and

$$\{\beta, \gamma\} = U(\gamma) \setminus \{U((2, 0, 1)) \cup U((2, 1, 1)) \cup U(\theta) \cup U(\lambda)\}.$$

Thus we have $c(\{\beta, \gamma\}) = c(\{\gamma\}) = 1$. One can verify also that $c(\{\alpha, \gamma\}) = 2$.

We have also $c(\{\gamma, \theta, \lambda\}) = 2$ and one can verify that

$$\begin{aligned}\{\gamma, \theta, \lambda\} &= [\gamma; \beta, \lambda] \cup [\lambda; (1,2,2), (1,0,3)] \\ &= [\gamma; \beta, (1,0,3)] \cup [\theta; (1,2,2), (1,1,3)].\end{aligned}$$

We use propositional logics in finding the blocks of a given set of MBs. It is well known in propositional logics that all logical formulas can be converted into *full disjunctive normal form* (DNF). More exactly, if α is a constraint of items (which is namely a logical formula), then by using simple transformations we can find the full DNF of α

$$\alpha = \bigvee_{i=1}^{n} \left[\bigwedge_{k=1}^{m_i} \beta_k^i \wedge \bigwedge_{k=1}^{n_i} (\neg \gamma_k^i) \right],$$

where $\beta_k^i, \gamma_k^i \in \Omega$, β_k^i, γ_k^i appear in α. One can verify that

$$U\left(\left[\bigwedge_{k=1}^{m_i} \beta_k^i \wedge \bigwedge_{k=1}^{n_i} (\neg \gamma_k^i) \right] \right) = [\beta_1^i, \ldots, \beta_{m_i}^i | \gamma_1^i, \ldots, \gamma_{n_i}^i]$$

is a block. By this in fact we have proved the following

Theorem 5.3. *(Finding full customer blocks of MBs.)*

(1) *There is an algorithm by that for any constraint of MBs α we can find the system of full customer blocks of $U(\alpha)$, i.e. we can find*

$$\{[\alpha_1^i, \alpha_2^i, \ldots, \alpha_{m_i}^i | \beta_1^i, \beta_2^i, \ldots, \beta_{n_i}^i] | i = 1, 2, \ldots, n\}$$

where $\alpha_1^i, \alpha_2^i, \ldots, \alpha_{m_i}^i, \beta_1^i, \beta_2^i, \ldots, \beta_{n_i}^i$ are all MBs that appear in α, such that

$$U(\alpha) = \bigcup_{i=1}^{k} [\alpha_1^i, \alpha_2^i, \ldots, \alpha_{m_i}^i | \beta_1^i, \beta_2^i, \ldots, \beta_{n_i}^i].$$

(2) *The decomposition of $U(\alpha)$ into full customer blocks is unique.*

(3) *The minimal representations of $U(\alpha)$ can be obtained from decomposition of $U(\alpha)$ into full customer blocks by combining some full customer blocks into one to reduce the number of blocks.*

(4) *The complexity of $U(\alpha)$ does not exceed the number of full clauses in the full DNF of α.*

Proof.

1. The well known algorithm in propositional logics converts a constraint of MBs α into full DNF. By this algorithm we can find the system of full customer blocks of $U(\alpha)$.

2. This is a result in propositional logics.

3. If

$$U(\alpha) = \bigcup_{i=1}^{k} [\alpha_1^i, \alpha_2^i, \ldots, \alpha_{m_i}^i | \beta_1^i, \beta_2^i, \ldots, \beta_{n_i}^i]$$

is a minimal representations of $U(\alpha)$ where, for example, some block $[\alpha_1^i, \alpha_2^i, \ldots, \alpha_{m_i}^i | \beta_1^i, \beta_2^i, \ldots, \beta_{n_i}^i]$ is not full. Then using the equivalence $X \equiv (X \wedge a) \vee (X \wedge \neg a)$ we can insert into the block the missing item a. In result we have the decomposition of $U(\alpha)$ into full customer blocks, which, accordingly to 2., is unique. The reverse transformation converts the full DNF of α into the given minimal representation of $U(\alpha)$.

4. The proof is evident. ■

Let us consider an example.

Example 5.4. *Following the Example 4.2 let a = "Flour", b = "Egg", c = "Milk", which can be identified by $a = (1,0,0)$, $b = (0,1,0)$, and $c = (0,0,1)$, respectively. The constrain $\alpha = (a \wedge b \to c)(\neg b \to (a \vee c))$ characterises the set of all those customers, who if buy flour and egg then buy also milk, and if do not buy egg, then would buy flour or milk. Let us denote this set of customers by A, i.e. $A = U(\alpha)$. By using simple transformations we have the full DNF of α:*

$$\alpha = (a \wedge b \wedge c) \vee (\neg a \wedge b \wedge c) \vee (\neg a \wedge b \wedge \neg c) \vee (\neg a \wedge \neg b \wedge c) \vee (a \wedge \neg b \wedge c) \vee (a \wedge \neg b \wedge \neg c).$$

The full customer block of $A = U(\alpha)$ is

$$A = U(\alpha) = [a, b, c] \cup [b, c|a] \cup [a, b|c] \cup [c|a, b] \cup [a, c|b] \cup [a|b, c].$$

One can remark that

$$\alpha = c \vee (\neg a \wedge b) \vee (a \wedge \neg b).$$

Thus one of the minimal representations of $A = U(\alpha)$ is

$$A = U(\alpha) = [c|] \cup [a|b] \cup [b|a].$$

This means that A can be characterized as the union of three blocks of customers: the first block contains those customers who buy milk, the second block contains all customers who buy flour but do not buy eggs, and the third one is the block of all customers who buy eggs but do not buy flour. One can see that the complexity of A is 3 and the structure of A is clear.

6. Classification

The classification is an important topics in economy and other areas. It has been discussed in a wide range of studies and the sufficient summaries can be found in extensive overviews of the theme ([3]). A multi-factor customer classification evaluation model was proposed in [9]. Some other problems arising in the classification on multiple database relations were considered in [10]. In general the following problems should be solved in the classification processes:

1. Determination of the characteristics of the objects to be classified. The objects may be items, transactions or customers. Finding the suitable representation for the objects is one of the most important tasks: an appropriate representation of the object's characteristics makes the model more simple and clearer that facilitates the more efficient classification algorithms and therefore yields more exact results.

2. Creating the classification algorithms that solve the problems in different areas and analysing the efficiency of these algorithms. The natures of the problems arising in different application areas are quite different, therefore most of the solutions of them are based particularly on the specificities of the areas.

3. Accordingly to the representation and classification processes the evaluation methods vary in the wide range of applications.

In the same formalism defined in the previous sections an approach to the study of customer's (or market basket's) classification is proposed here, based on the quantities of the items purchased by the customers, or on the quantities of the items involved in the transactions, respectively.

Classifications: The concept of customer classification can be generalized. Let Ω be an arbitrary set of customers (or transactions) and $A \subseteq \Omega$. Then a *classification* of A is a family of subsets of A, $S = \langle U_1, U_2, \ldots, U_k \rangle$, where $U_i \subseteq A$ for all $i = 1, 2, \ldots, k$. The set of all classifications of a given set A is denoted by $CLASS(A)$. A classification is *total classification* if it covers A, i.e. $\bigcup_{i=1}^{k} U_i = A$. In the following we deal mainly with total classifications. A total classification is called a *partition* of A if the blocks are pairwise disjoint, i.e. $U_i \cap U_j = \emptyset$ for all $i \neq j$.

By definitions one can see that two natural orders are usually considered on the set of classifications. The first one is the inclusion that holds for two classifications S, Q, if $S \subseteq Q$. The second order is defined as the fineness of the classifications: for two classifications S, Q, $S = \langle U_1, U_2, \ldots, U_k \rangle$, $Q = \langle V_1, U_2, \ldots, V_l \rangle$ we write $S \leq Q$ if for all $i = 1, \ldots, k$ there exists $1 \leq j \leq l$ such

that $U_i \subseteq V_j$.

Operations on the classifications: The following operations on the classifications should be considered:

1. *0-level operations (Valuations):* These are the operations that associate each classification to a number: $F : CLASS(A) \to \mathbb{R}$. For $S \in CLASS(A)$ then $F(S)$ is considered as a kind of valuation or indicator of S. For a classification $S = \langle U_1, U_2, \ldots, U_k \rangle$ the following valuations are often typically considered:

(1) $F(S) = |S|$: $F(S)$ denotes the number of classes in the classification. For this valuation there are two trivial classifications $S = \langle A \rangle$ and $S = \langle \{a\} | a \in A \rangle$. In these cases $F(S) = 1$ and $F(S) = |A|$, respectively.

(2) $F(S) = max\{|U_i|\}$: $F(S)$ denotes the maximal number of customers in the classes of the classification.

(3) $F(S) = \frac{1}{k} \sum_{i=1}^{k} |U_i|$: $F(S)$ denotes the average number of customers in the classes of the classification.

(4) $F(S) = \sum_{i=1}^{k} \lambda(U_i)|U_i|$ where $\lambda : 2^{\Omega} \to \mathbb{R}$ is an evaluation that assigns to each $U \subseteq \Omega$ a value $\lambda(U) \in \mathbb{R}$. If $\lambda(U)$ denotes the tariff posed on each customer in the U class, then $F(S)$ is the total tariff obtained from the customers.

A typical problem related to the valuations of classifications is as follows: Let A be a given set of customers, F_1, F_2 be two valuations of the classifications on A and let M be a given limit. Then the problem is to find $S \in CLASS(A)$ such that

(i) $F_1(S) \leq M$, and

(ii) $F_2(S) \to max$.

The other optimal problems may be formulated in similar way.

2. *1-level operations (Selections):* These are the operations that based on the given classification select out a class (or classes) of customers: $F : CLASS(A) \to 2^A$, or $F : CLASS(A) \to 2^{2^A}$, where 2^A denotes the family of all subsets of A. A typical selection is the representative selection, where $F : CLASS(A) \to 2^A$ such that for each $S = \langle U_1, U_2, \ldots, U_k \rangle \in CLASS(A)$ we have

(i) $|F(S) \cap U_i| = 1$ for all $1 \leq i \leq k$, and

(ii) $F(S) \subseteq \bigcup_{i=1}^{k} U_i$.

Another familiar example of selection is tariff-based selection: If $\lambda : 2^A \to \mathbb{R}$ is an evaluation and M is a threshold, then $F_{\lambda,M} : CLASS(A) \to 2^A$ where $F_{\lambda,M}(S) = \{U_i \in S | \lambda(U_i) \leq M\}$.

3. *2-level operations (Transformations):* These are the operations that transform a given classification into another one, $F : CLASS(A) \to CLASS(A)$. For a classification $S = \langle U_1, U_2, \ldots, U_k \rangle$ and for a given $C \subseteq A$:

(1) *Restriction:* Let $F_C(S) = \langle U_1 \cap C, U_2 \cap C, \ldots, U_k \cap C \rangle$. $F_C(S)$ is a restriction of S.

(2) *Extension:* Let $F_C(S) = \langle U_1 \cup C, U_2 \cup C, \ldots, U_k \cup C \rangle$. $F_C(S)$ is an extension of S.

(3) *Multiplication:* If $S = \langle U_1, U_2, \ldots, U_k \rangle$, $Q = \langle V_1, V_2, \ldots, V_l \rangle$, then $S \times Q = \langle U_i \cap V_j | i = 1, \ldots, k, j = 1, \ldots, l \rangle$.

(4) *Exponentiation:* Let $S^1 = S$ and $S^{m+1} = S^m \times S$, or $S^m = \langle U_{i_1} \cap \ldots \cap U_{i_m} | 1 \leq i_1 < \ldots < i_m \leq k \rangle$.

It should be noted that, in fact, the selections can be considered as special transformations. The problems of classifications are often set up with the valuations, the selections and the transformations.

Compactness and efficiency of a classification. As a customer (or a transaction) is a tuple $\alpha = (\alpha[1], \alpha[2], \ldots, \alpha[n])$, where $\alpha[i] \in \mathbb{N}$ is the quantity of item p_i in α, different metric can be defined between customers. One of these is the Euclidean metric: the metric between the two customers $\alpha = (\alpha[1], \alpha[2], \ldots, \alpha[n])$, $\beta(\beta[1], \beta[2], \ldots, \beta[n])$ is

$$d(\alpha, \beta) = \left[\sum_{i=1}^{n} (\alpha[i] - \beta[i])^2 \right]^{\frac{1}{2}}.$$

The metric between customers may be understood as a kind of similarity between customers and the choice of suitable metric on the set of customers is one of the significant factors that determines the efficiency of the classification process.

Let $d(\alpha, \beta)$ denote the distance between two customers α, β and B be a set of customers. We say that a number $r \in \mathbb{N}$ is the *radius* of B if

(i) There exists $\alpha \in B$ such that for all $\beta \in B$ we have $d(\alpha, \beta) \leq r$, and

(ii) r is the smallest number that satisfies i.

Then we say also that α is a *center* of B. A finite set of customers has unique radius, but may have more than one center. The radius of a set of customers B is denoted by $r(B)$.

For a classification $S = \langle U_1, U_2, \ldots, U_k \rangle$ the *compactness* of S is determined by two factors: the number of classes in S, namely $|S|$, and the radius of the classes in S. For $n, m \in \mathbb{N}$ we say that a classification $S = \langle U_1, U_2, \ldots, U_k \rangle$ is $(n, m)-$*compact*, if

(i) $|S| \leq n$, and

(ii) $r(U_i) \leq m$ for all $i = 1, \ldots, k$ and

(iii) there is no $m_1 < m$ that satisfies ii.

The criterions for compactness of classifications should be defined by the experts of the application areas. In practice many optimal problems are posed on the set of classifications with given compactness. A company may require a classification of customers that, with some bound on the number of customer classes as well as on the sizes of the classes, provides maximal revenue.

Neighborhood of the customers. Based on the concept of distance between two customers, the neighborhood may be formulated consequently. Let $d(\alpha, \beta)$ denote the distance between two customers α, β and $m \in \mathbb{N}$ then we say that β is a $m-$*neighbor* of α if $d(\alpha, \beta) \leq m$. The *nearest neighbor* hence may be determined in similar way: β is a *nearest neighbor* of α if β is a $m-$neighbor of α, and α has no other $p-$neighbor, where $p < m$.

Another method to define the neighborhood of the customers may be as follows: Let A be a set of customers and $S = \langle U_1, U_2, \ldots, U_k \rangle$ be a classification on A, then we say that β is a $m-$*neighbor* of α (in S) if there exist $\{U_{i_1}, U_{i_2}, \ldots, U_{i_m}\} \subseteq S$ such that $\alpha, \beta \in \bigcap_{j=1}^{m} U_{i_j}$. The *rank* of the neighborhood between α, β is the greatest m such that β is $m-$neighbor of α. The *nearest neighbor* of α in this sense is those β that is $m-$neighbor of α and α has no other neighbor with really higher rank of neighborhood. By using the above notion of exponents we can see that α, β are $m-$neighbors if and only if there exists $V \in S^{(m)}$ such that $\alpha, \beta \in V$.

For $S = \langle U_1, U_2, \ldots, U_k \rangle$ let R_S be a relation on A such that

$$(\alpha, \beta) \in R_S \Longleftrightarrow \exists i : \alpha, \beta \in U_i.$$

One can verify that if S is total classification, then R_S is a reflexive and symmetric relation on A.

We should note that β is an $m-$neighbor of α if and only if $\alpha, \beta \in V$, for some $V \in S^{(m)}$, i.e. if and only if $(\alpha, \beta) \in R_{S^{(m)}}$. We have

Lemma 6.1. *$R_{S^{(m)}}$ is the relation of $m-$neighborhood induced by S, i.e. β is an $m-$neighbor of α if and only if $(\alpha, \beta) \in R_{S^{(m)}}$.*

We should recall also that $S \leq Q$ denotes the fineness order between S and Q: we write $S \leq Q$ if for all $U_i \in S$ there exists $V_j \in Q$ such that $U_i \subseteq V_j$. It is easy also to see that the following lemma holds.

Lemma 6.2. *If S, Q are two classifications such that*

(i) *$Q \leq S$, and*

(ii) *$S \leq Q$,*

then $R_S = R_Q$.

For a classification $S = \langle U_1, U_2, \ldots, U_k \rangle$ let

$$S^{max} = \{U_i \in S | \nexists U_j \in S, i \neq j : U_i \subsetneq U_j\}.$$

By Lemma 6.2. we have $R_S = R_{S^{max}}$.

As a consequence of Lemma 6.1, 6.2 we can construct an algorithm that for a given classification S efficiently yields $(S^{(m)})^{max}$ by which we can easily determine the $m-$neighborhood relation, and therefore, the nearest neighborhood relation between customers or transactions.

Algorithm 6.1.

Input: A classification $S = \langle U_1, U_2, \ldots, U_k \rangle$, $m \in \mathbb{N}$.

Output: The classification $Q = (S^{(m)})^{max}$.

Step 1: Compute

$$S^{(m)} = \langle U_{i_1} \cap U_{i_2} \cap \ldots \cap U_{i_m} | 1 \leq i_1 < i_2 < \ldots < i_m \leq k \rangle.$$

Step 2: Compute $Q := (S^{(m)})^{max}$.

Let us consider the inverse problem: Based on a relation between the customers how can we construct a customer classification such that the customers in the same class are in relation each with other? Let A be a set of customers, $U \subseteq A$ and $R \subseteq A \times A$ be a reflexive relation on A. We say that U is a complete block of R if for all $\alpha, \beta \in U$ we have $(\alpha, \beta) \in R$.

Lemma 6.3. *Let A be a set of customers. Then for all reflexive relations $R \subseteq A \times A$ there exists a classification $S = \langle U_1, U_2, \ldots, U_k \rangle$ on A such that*

(i) *U_i is a complete block of R for all $1 \leq i \leq k$, and*

(ii) *there is no other classification S' of A that satisfies 1. and $S \leq S'$, where $S \leq S'$ is the order between classifications.*

The decomposition of a reflexive relation into complete blocks is not unique. The following simple procedure produces for the given $\alpha \in A$ a complete block $C(\alpha)$ that contains α : Suppose that $A = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$, then

(1) $i := 0$, $n := 0$, $C^i(\alpha) := \{\alpha\}$,

(2) If $i + 1 \leq m$ then

$i := i + 1$

If there exists k, $n + 1 \leq k \leq m$ such that

$$\forall \gamma \in C^{i-1}(\alpha) : (\alpha_k, \gamma), (\gamma, \alpha_k) \in R$$

then

$n :=$ the smallest such k,

$C^i(\alpha) := C^{i-1}(\alpha) \cup \{\alpha_n\}$.

Go back to (2).

Else go to (3).

Else go to (3).

(3) $C(\alpha) := C^i(\alpha)$

Stop.

For a given $\alpha \in A$ the complete block $C(\alpha)$ given by the procedure is not unique. The following algorithm produces for a given set of customers a classification that consists of complete blocks. Let $A = \{a_1, a_2, \ldots, a_p\}$ and R be a reflexive relation on A.

Algorithm 6.2.

Input: A relation $R \subseteq A \times A$.

Output: A classification $S = \langle U_1, U_2, \ldots, U_k \rangle$, where U_i's are complete blocks of R.

Step 1: $i := 1$; $j := 1$; $b_i := a_1$;

Step 2: By the above procedure with few modifications let us compute $C(b_i)$:

(1) $C^1(b_i) := \{b_i\}$, and

(2) For $s = 1, 2, \ldots$ let
$C^{s+1}(b_i) := C^s(b_i) \cup \{\beta \in A | \beta \notin \bigcup_{r=1}^{i-1} C(b_r) \wedge \forall \gamma \in C^s(b_i) : (\beta, \gamma), (\gamma, \beta) \in R\}$.

(3) $C(b_i) := C^s(b_i)$ if $C^{s+1}(b_i) = C^s(b_i)$.

If $j < p$, then $i := i+1$. Let m be the smallest index such that $j < m < p$ and $a_m \notin \bigcup_{l=1}^{i-1} C(b_l)$. Put $j := m$ and $b_i := a_m$. Return to Step 2.

Otherwise, if $j = p$, then stop.

The algorithm may give different classifications. The following example illustrates the above discussions of classification and algorithm.

Example 6.1. *Let $\{p_1, p_2, p_3, p_4, p_5\}$ be the set of items and $A = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$ be the set of customers whose purchases are shown in the table below.*

Table 1. Customer's purchases

Customers	p_1	p_2	p_3	p_4	p_5
α_1	*3*	*0*	*1*	*0*	*0*
α_2	*5*	*1*	*0*	*1*	*1*
α_3	*1*	*5*	*6*	*0*	*1*
α_4	*0*	*6*	*7*	*1*	*1*
α_5	*1*	*4*	*8*	*2*	*6*
α_6	*2*	*0*	*6*	*6*	*7*
α_7	*0*	*0*	*0*	*5*	*7*

Let $S = \langle U_1, U_2, \ldots, U_5 \rangle$ be the classification where U_i denotes the set of those customers who buy (some of) p_i item, $i = 1, 2, \ldots, 5$:

$U_1 = \{\alpha_1, \alpha_2, \alpha_3, \alpha_5, \alpha_6\}$, $U_2 = \{\alpha_2, \alpha_3, \alpha_4, \alpha_5\}$,
$U_3 = \{\alpha_1, \alpha_3, \alpha_4, \alpha_5, \alpha_6\}$, $U_4 = \{\alpha_2, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$,
$U_5 = \{\alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$.

By Algorithm 6.1 we can compute

Table 2. The classification with $m-$neighborhood classes

m	$(S^{(m)})^{max}$
1	$\{\alpha_1, \alpha_2, \alpha_3, \alpha_5, \alpha_6\}$, $\{\alpha_1, \alpha_3, \alpha_4, \alpha_5, \alpha_6\}$, $\{\alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$
2	$\{\alpha_1, \alpha_3, \alpha_5, \alpha_6\}$, $\{\alpha_2, \alpha_3, \alpha_5, \alpha_6\}$, $\{\alpha_2, \alpha_3, \alpha_4, \alpha_5\}$, $\{\alpha_3, \alpha_4, \alpha_5, \alpha_6\}$, $\{\alpha_2, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$
3	$\{\alpha_2, \alpha_3, \alpha_5\}$, $\{\alpha_3, \alpha_5, \alpha_6\}$, $\{\alpha_3, \alpha_4, \alpha_5\}$, $\{\alpha_2, \alpha_5, \alpha_6\}$, $\{\alpha_2, \alpha_4, \alpha_5\}$, $\{\alpha_4, \alpha_5\alpha_6\}$
4	$\{\alpha_3, \alpha_5\}$, $\{\alpha_2, \alpha_5\}$, $\{\alpha_5, \alpha_6\}$, $\{\alpha_4, \alpha_5\}$

The rank of the neighborhood between the customers is shown in the following table:

Table 3. The rank of the neighborhood between customers.

	α_1	α_2	α_3	α_4	α_5	α_6	α_7
α_1	5	1	2	1	2	2	0
α_2	1	5	3	3	4	3	2
α_3	2	3	5	3	4	3	1
α_4	1	3	3	5	4	3	2
α_5	2	4	4	4	5	4	2
α_6	2	3	3	3	4	5	2
α_7	0	2	1	2	2	2	5

Thus for a given $k, 0 \leq k \leq 5$, two customers α_i, α_j are $k-$neighbors if α_i, α_j buy at least k similar items. By removing from the above table all the neighborhoods of the rank less than k we obtain a relation on A. The application of the Agorithm 2 will yield a complete classification. The classifications for $k = 2, 4, 5$ are presented in the following table.

Table 4. The classification induced by $k-$neighborhood between customers.

Rank	Classification
$k = 2$	$\langle\{\alpha_1, \alpha_3, \alpha_5, \alpha_6\}, \{\alpha_2, \alpha_4, \alpha_7\}\rangle$
$k = 4$	$\langle\{\alpha_1\}, \{\alpha_2, \alpha_5\}, \{\alpha_3\}, \{\alpha_4\}, \{\alpha_6\}, \{\alpha_7\}\rangle$
$k = 5$	$\langle\{\alpha_1\}, \{\alpha_2\}, \{\alpha_3\}, \{\alpha_4\}, \{\alpha_5\}, \{\alpha_6\}, \{\alpha_7\}\rangle$

7. Conclusion

The paper overviews the results of some previous researches concerning the customer's market baskets and proposes a generalization of the concept of customer classification. In the formalism presented here and in previous researches the market baskets, the customers, or the transactions are studied in more details by their quantity involved in the transactions. The first advantage of the approach is that the market baskets, the customers, or the transactions are characterized as sets of quantities of items. This implies that the market baskets, the customers, or the transactions though having different roles and meaning in different application areas can be studied in a unique form as sets of quantities of items. Secondly, the formalism reveals the natural structure of the market baskets (therefore, of the customers, or of the transactions). Based on this structure the frequent market baskets, the association rules between them, the constraints and the complexity of customers, as well as the classification of customers are studied. The results of the previous researches and of this study show that the formalism offers efficient methods for analysing the problems of market baskets and customers.

The formalism discussed in this paper reveals also a new aspect for further studies. One can remark that in this paper only the natural structure of market baskets is dealt with. However in different application areas beside this natural structure the market baskets possess also other particular structures that are imposed intentionally or unintentionally. Thus the market baskets and customers should be studied in the more complex structures that cover both the natural structure and the particular structures. This may be interesting topics for further studies.

References

[1] **Agrawal, R. and R. Srikant,** Fast algorithms for mining association rules, in: *Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994,* 487–499.

[2] **Benczúr, A. and Gy.I. Szabó,** Functional dependencies on extended relations defined by regular languages, *Foundations of Information and Knowledge Systems, Kiel, Germany, 2012,* eds. T. Lukasiewicz and A. Sali, LNCS **7153**, 384–404.

[3] **Chicco G., Napoli R., Piglione F., Postolache P., Scutariu M., Toader C.,** Emergent Customer Classification, Generation, Transmission and Distribution, *IEE Proceedings*, **152**, 2, 2005, 164–172.

[4] **Demetrovics, J., Hua Nam Son and A. Guban,** An algebraic representation of frequent market baskets and association rules, *Cybernetics and Information Technologies*, **11** (2) (2011), 24–31.

[5] **Demetrovics, J., G.O.H. Katona, D.M. Miklós and B. Thalheim,** On the number of independent functional dependencies, LNCS **3861**, 2006, 83–91.

[6] **Mannila, H. and H. Toivonen,** Discovering generalized episodes using minimal occurrences, *Proc. of the Second Int. Conf. on Knowledge Discovery and Data Mining (KDD' 96),* AAAI Press, 1996, 146–151.

[7] **Pasquier, N., Y. Bastide, R. Taouil and L. Lakhal,** Discovering frequent closed itemsets for association rules, *Proc. of the 7th Int. Conf. on Database Theory, ICDT'99, London, 1999,* 398–416.

[8] **Ping-Yu Hsu, Yen-Liang Chen and Chun-Ching Ling,** Algorithms for mining association rules in bag databases, *Information Sciences*, **166** (1-4) (2004), 31–47.

[9] **Qiaohong Zu, Ting Wu and Hui Wang,** A multi-factor customer classification evaluation model, *Computing and Informatics*, **29**, 2010, 509–520.

[10] **Thangaraj, M. and C.R.Vijayalakshmi,** A study on classification approaches across multiple database relations, *Int. J. of Computer Applications, 0975-8887,* **12** (12) (2011), 1–6.

J. Demetrovics
MTA SZTAKI
Budapest, Hungary
demetrovics@sztaki.hu

Hua Nam Son
Budapest Business School
Budapest, Hungary
huanamson@yahoo.com

A. Guban
Budapest Business School
Budapest, Hungary
guban.akos@pszfb.bgf.hu