Hua Nam Son[1]

The market basket model In information-deficient environment

LOGISZTIKA – INFORMATIKA – MENEDZSMENT
volume 3 ● number 1 ● március 2018 ● pp: 49-59
DOI: 10.29177/LIM.2018.1.49

Abstract

In this research the market baskets are is studied in the environment with insufficient information. In the environment with sufficient information the frequency of items and market baskets, the confidency of the association rules can be determined exactly, by which the management strategies can be decided definitely. The problems become more difficult in the environment with insufficient information. The managers in this case have to cope with information deficience and have to decide the management strategies based on insufficient information. In this paper the concepts of frequent items and frequent market baskets, the association rules and their confidency are formulated in a more generalized model with empty data. Because of the deficiency of information the frequency of market baskets and the confidency of association rules can be approximately appreciated only. The problem of determination of the information deficient items and extracting association rules in information deficient environments are shown to be interesting problems. The algorithms to determine the information deficience of items and market baskets, as well as to generate the all information-deficient items are presented also in this paper.

Keywords: market basket, frequent item, association rule, information deficient items.

[1] PhD, egyetemi docens, Budapesti Gazdasági Egyetem Pénzügyi és Számviteli Kar

1. Introduction

The extraction of hidden information from large databases is one of the goals of data mining. To solve the problems of economy, great efforts have been made to discover the informations hidden in the customer transactions (see, for examples [1], [2], [6]). As shown in [3], the market baskets (MB) model provide a research frame and methods for discovering the frequent items and the association rules between items. Therefore the market baskets model has important role in various applications, esspecially in decision making and strategy determination of retail economy.

As discussed in [3], in the previous studies of market baskets model the considerations are focussed on items and the associations between items in transactions. Then the researchers examine, for examples, bread, milk,. itself, but are not interested in the quantity of these items in transactions. Although, one can easily see that in practice, the quantity of items is one of the main factors that decides the business feature of the items. We know that 40% of customers buy eggs, but we know also 35% of customers buy 30 pieces of eggs, while only 0,1% of customers buy 100 pieces of eggs. The same question is for association rules. Therefore the quantitative analysis of transactions is necessary.

In [3] a model was set up that is suitable for quantitative analysises of transactions. In this model the researchers are concerned not with *itemsets* (see [1]), but infact, with *market baskets* or *transactions,* that contain items with concrete quantity.

The quantitative analysis reveals more informations hidden in the transactions. It turns out that, though there is an association between milk and butter, this associative relation can be proved in proportion of 2 liter milk and 0,5 kg butter, and this associative relation does not hold between 1 liter milk and 50 kg butter.

The new model provides new tools for research. The customer's market baskets, i.e. the transactions, in this model constitute a lattice with natural partial order. So the lattice-theoretic methods can be applied for transaction's examinations.

The problem seems to be more complicated when the researchers have to cope with the transactions with deficient information. Maybe, for some reason, the managers have to solve the problems with deficient data of transactions. Naturally, the similar questions can be raised for new model: how the managers can determine the frequent items and the associative relation between items with information deficiency. Nevertheless, besides these questions, it is interesting also to determine the information-deficient items. Identification of information-deficient items is certainly a greate support for business systems management.

In the following Sections 2-4 the main frame of the market baskets model as proposed in [3] and the important reasults are recalled. In Sections 5 we suggest a model in wich we can study the market baskets in information deficient environments.

2. Market Basket Model

Market Baskets

For a finite set of items $P = \{p_1, p_2, \dots, p_n\}$ we consider a market basket (MB) as a vector $\alpha = (\alpha[1], \alpha[2], \dots, \alpha[n])$, where $\alpha[i] \in \mathbb{N}$ is the quantity of p_i in the basket α, $\mathbb{N}$ is the set of all natural numbers. Let Ω denote the set of all MBs over P.

For $\alpha = (\alpha[1], \alpha[2], \dots, \alpha[n])$, $\beta = (\beta[1], \beta[2], \dots, \beta[n]) \in \Omega$ we write $\alpha \leq \beta$ if for all $i \leq n$ we have $\alpha[i] \leq \beta[i]$. $(\Omega, \leq)$ is a lattice with the natural partial order $\leq$. For a set $A \subseteq \Omega$ we denote $U(A) = \{\alpha \in \Omega | \forall\, \beta \in A: \beta \leq \alpha\}$ and $V(A) = \{\alpha \in \Omega | \forall\, \beta \in A: \alpha \leq \beta\}$. Let $sup(A) = \{\alpha \in U(A) | \nexists\, \beta \in U(A): \beta < \alpha\}$ and $inf(A) = \{\alpha \in V(A) | \nexists\, \beta \in V(A): \alpha < \beta\}$.

Support of Market Baskets

For a set $A \subseteq \Omega$ and $\alpha \in \Omega$ the support of α in A is denoted by

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

In words, $supp_A(\alpha)$ denotes the ratio of all market baskets that exceeds the given threshold α to the whole A. $supp_A(\alpha)$ of an market basket α is a statistical index that characterizes the support of α. The managers naturally, have to deal with special attention to those market baskets of high support, or, in contrary, to those market baskets of extremely low support.

Association between Market Baskets

In this market baskets model an item p_i should be identified with $U(\alpha_i)$, where $\alpha_i = (\alpha_i[1], \alpha_i[2], \dots, \alpha_i[n])$, $\alpha_i[k] = 0$ if $k \neq i$ and $\alpha_i[i] = 1$.

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

An *association rule* of β to α is denoted by $\alpha \rightarrow \beta$. The *confidence* of $\alpha \rightarrow \beta$ (in A) is the ratio

$$conf(\alpha \rightarrow \beta) = \frac{supp_A(\alpha \cup \beta)}{supp_A(\alpha)}$$

In words, $conf(\alpha \rightarrow \beta)$ is the ratio of all market baskets that support both α, β to those that support α only.

3. Frequent Market Baskets

Frequent items play important role in economic management, and therefore in data management (see, for examples, [7]). For a set $A \subseteq \Omega$, $\alpha \in \Omega$ and $0 \leq \varepsilon \leq 1$ we say that α is ε–frequent, if $supp_A(\alpha) \geq \varepsilon$. The set of all ε–frequent MBs is denoted by Φ_A^{ε}. Then the *Apriori Principle* now can be stated as followings:

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

Example 1: Let $P = \{a, b, c\}$ be a set of items and let $A = \{\alpha, \beta, \gamma, \delta\}$ be a set of transactions, where $\alpha = (2,1,0)$, $\beta = (1,1,1)$, $\gamma = (1,0,1)$, $\delta = (2,2,0)$. 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:

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

Let

$$\Phi_{A,k} = \{\alpha \in \Omega | \exists \alpha_1, \alpha_2, \dots, \alpha_k \in A: \alpha \leq \{\alpha_1, \alpha_2, \dots, \alpha_k\}\}$$

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

In [3] the following Theorem 1 was proved:

Theorem 1: For a finite set of items $P = \{p_1, p_2, \dots, p_n\}$, 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 $\alpha_1, \alpha_2, \dots, \alpha_k \in A$ such that $\alpha \in L(\alpha_1, \alpha_2, \dots, \alpha_k)$ where $k = \lceil \varepsilon |A| \rceil$.

Based on Theorem 1 an algorithm was proposed in [3] to produce the set of all ε-frequent MBs for a given set of transactions $A \subseteq \Omega$:

Theorem 2: For a finite set of items $P = \{p_1, p_2, \dots, p_n\}$ there exists an algorithm that for a set of MBs $A \subseteq \Omega$ and a threshold $0 \leq \varepsilon \leq 1$ produces the set of all ε–frequent MBs Φ_A^ε.

As a direct consequence of the previous theorems we have:

Theorem 3: (Explicit representation of *large MBs*) For a set of items $P = \{p_1, p_2, \dots, p_n\}$, a set of MBs $A \subseteq \Omega$ and a threshold $0 \leq \varepsilon \leq 1$ there exist $\alpha_1, \alpha_2, \dots, \alpha_s \in \Omega$ where $s = \binom{|A|}{\lceil \varepsilon |A| \rceil}$ such that

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

In fact for a given A we can choose $\alpha_1, \alpha_2, \dots, \alpha_s \in \Omega$ that is in a sense the 'smallest' set of MBs generating Φ_A^ε. Such $\alpha_1, \alpha_2, \dots, \alpha_s$ that satisfies:

i. $\Phi_A^\varepsilon = \bigcup_{i=1}^{s} L(\alpha_i)$
ii. $\forall i, j \leq s$: $(\alpha_i \nleq \alpha_j) \wedge (\alpha_j \nleq \alpha_i)$.

is considered as a basic set of MBs that generates Φ_A^ε. As shown in [3] we have really:

Theorem 4:
1. For a set of items P, a threshold $0 \leq \varepsilon \leq 1$ every set of MBs $A \subseteq \Omega$ has an unique basic ε- frequent set of MBs S_A^ε.

2. There is an algorithm that creates the unique basic ε- frequent set of MBs for a given set of MBs $A \subseteq \Omega$ and a given threshold ε.

For a set of items P, a set of MBs $A \subseteq \Omega$ and a threshold $0 \leq \varepsilon \leq 1$ if $|P| = n, k = \lceil \varepsilon |A| \rceil$ and $m = max\{\alpha[i] |\ i = 1,2, \dots, n, \alpha \in A\}$ then it is easy to see that $|\underset{A}{S^s}| \leq \binom{|A|}{k}$ and therefore there is an algorithm that produces $\underset{A}{S^s}$ in $O\ (\binom{|A|}{k}. m. n)$ time.

4. Association of Market Baskets and the Confidence of Association Rules

Like determination of frequent items exploring the associations between items is an important problem in economic management and in data management (see, for examples, [4], [7], [8]). By the association we mean the hidden relations between items that are revealed by the behavior of the customers. In this generalized model we can study the association between MBs. In words we say that there is an associative relation between two MBs α, β, in a certain group of customers A, if most of customers in A who buy α also buy β. More exactly, we denote the association of β from α by $\alpha \to \beta$. The confidence of the rule $\alpha \to \beta$ is

$$\underset{A}{conf}\ (\alpha \to \beta) = \frac{supp_A(\alpha \cup \beta\ )}{supp_A(\alpha)}.$$

For a set of items P, a set of MBs $A \subseteq \Omega$ and a threshold $0 \leq \varepsilon \leq 1$ an association $\alpha \to \beta$ is ε-confident if $conf_A(\alpha \to \beta)$. The set of all ε-confident associations of A is denoted by C_A^s .

In [5] a condition for ε-confident association rules, an explicit representation of association rules were given and based on that we have an algorithm to find all ε -confident association rules for given left side.

5. Market Basket Model in Information Deficient Envirement

In this Section we propose a generalization of the market basket model. In an Information Deficient Environment (IDE) the managers may have insufficient information about items in market baskets. Two problems emerge in this case: How can we determine the frequent sets of MBs and the association rules if we possess inadequate information of items, and how can we point out those items of which we possess inadequate information.

Market Baskets

Let $\mathbb{N}^* = \mathbb{N} \cup \{*\}$ where $*$ denotes the null value. A $*$-market basket of the finite set of items $P = \{p_1, p_2, \dots, p_n\}$ is a vector $\alpha = (\alpha[1], \alpha[2], \dots, \alpha[n])$, where $\alpha[i] \in \mathbb{N}^*$. Thus $\alpha[i]$ is the quantity of p_i in the basket α, if $\alpha[i] \in \mathbb{N}$, or $\alpha[i]$ denotes the lack of information of p_i in the basket α, if $\alpha[i] = *$. We denote the set of all $*$-MBs over P by Ω^*.

For $\alpha = (\alpha[1], \alpha[2], \dots, \alpha[n]), \beta = (\beta[1], \beta[2], \dots, \beta[n]) \in \Omega^*$ we write

$\alpha \preccurlyeq_{min} \beta$ if for all $i \leq n$: $(\alpha[i] \in \mathbb{N}) \Rightarrow (\beta[i] \in \mathbb{N}) \wedge (\alpha[i] \leq \beta[i])$ and $\alpha[i] = \beta[i] = *$ otherwise.

$\alpha \preccurlyeq_{max} \beta$ if for all $i \leq n$: $(\alpha[i] \in \mathbb{N}) \Rightarrow (\beta[i] \in \mathbb{N}) \wedge (\alpha[i] \leq \beta[i])$.

One can remark that $(\Omega^*, \preccurlyeq_{min})$ is a poset with the partial order $\preccurlyeq_{min}$, i.e. $\preccurlyeq_{min}$ is a reflexive, antisymmetric, and transitive relation on Ω^*, while $\preccurlyeq_{max}$ is a reflexive, but not certainly antisymmetric or transitive relation.

Support

For a set $A \subseteq \Omega^*$ and $\beta \in \Omega^*$ the support of β in A can be defined now in different ways:
Let

$$supp_A^{min}(\beta) = \frac{|\{\alpha \in A |\ \beta \preccurlyeq_{min} \alpha\}|}{|A|}$$

$supp_A^*(\beta)$ denotes the ratio of all market baskets that exceeds β to the whole A.

We consider also

$$supp_A^{max}(\beta) = \frac{|\{\alpha \in A | \beta \preccurlyeq_{max} \alpha\ \}|}{|A|}$$

$supp_A^{max}(\beta)$ denotes the ratio of all market baskets that exceeds β at all information-definite items to the whole A.

$supp_A^{min}(\beta)$ and $supp_A^{max}(\beta)$ are statistical indicators that characterize the support of β in the set of market baskets A. One can remark that $supp_A^{min}(\beta) \leq supp_A^{max}(\beta)$, and both $supp_A^{min}(\beta)$, $supp_A^{max}(\beta)$ do not determine exactly the real support of β in A. Even so by $supp_A^{min}(\beta)$ and $supp_A^{max}(\beta)$ the managers can figure out approximately the support of β in A.

The concept of support can be generalized for sets of MBs. For a set $A \subseteq \Omega^*$ the minimal and maximal support of $B \subseteq \Omega^*$ is denoted by $supp_A^{min}(B)$ and $supp_A^{max}(B)$, respectively, where:

$$supp_A^{min}(B) = \frac{|\{\alpha \in A | \forall \beta \in B: \beta \preccurlyeq_{min} \alpha\}|}{|A|}$$

and

$$supp_A^{max}(B) = \frac{|\{\alpha \in A | \forall \beta \in B: \beta \preccurlyeq_{max} \alpha\}|}{|A|}.$$

Items	p_1 Milk (l)	p_2 Butter (gr)	p_3 Pacifier (piece)	p_4 Pampers (piece)	p_5 Hammer (piece)
α_1	1	200	*	0	0
α_2	2	400	1	2	0
α_3	1	200	1	*	0
α_4	0	0	0	*	1
α_5	6	1200	0	0	0

Table 1: A set of MBs with given set of items.

Example:
Let $A = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$ be a set of MBs as in Table 1 and $\alpha = (1,200,0,*,0)$ we have $supp_A^{min}(\alpha) = \frac{|\{\alpha_3\}|}{|A|} = \frac{1}{5}$ and $supp_A^{max}(\alpha) = \frac{|\{\alpha_3,\alpha_5\}|}{|A|} = \frac{2}{5}$.

Frequent items

Let $A \subseteq \Omega^*$ be a set of MBs and a threshold value $\varepsilon \geq 0$ we say that an MB α is strongly $\varepsilon -$frequent, or weakly $\varepsilon -$frequent in A if $supp_A^{min}(\alpha) \geq \varepsilon$, or $supp_A^{max}(\alpha) \geq \varepsilon$, respectively.

One can remark that if $A \subseteq \Omega^*$ be a set of MBs, $|A| = m$ over the set of items $P = \{p_1, p_2, \ldots, p_n\}$, $|P| = n$ then for any $\alpha \in \Omega^*$, $supp_A^{min}(\alpha)$ and $supp_A^{max}(\alpha)$ can be computed in $O(n \times m)$ steps. Thus in $O(n \times m)$ steps we can decide if an given MB is strongly $\varepsilon -$frequent or weakly $\varepsilon -$frequent for given threshold value $\varepsilon \geq 0$. In fact we have:

Theorem 5:
There is an algorithm which checks in $O(n \times m)$ time that for a set of MBs $A \subseteq \Omega^*$, $|A| = m$ over the set of items $P = \{p_1, p_2, \ldots, p_n\}$, $|P| = n$ if a given MB $\alpha \in \Omega^*$ is strongly $\varepsilon -$frequent or weakly $\varepsilon -$frequent for given threshold value $\varepsilon \geq 0$.

Example:
Let $A = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$ be a set of MBs as in Table 1 and $\alpha = (1,200,0,*,0)$ we have $supp_A^{min}(\alpha) = \frac{|\{\alpha_3\}|}{|A|} = \frac{1}{5}$ and $supp_A^{max}(\alpha) = \frac{|\{\alpha_3,\alpha_5\}|}{|A|} = \frac{2}{5}$. Thus for $\varepsilon = 30\%$ α is weakly $\varepsilon -$frequent, but is not strongly $\varepsilon -$frequent in A.

Association between Market Baskets

Let $A \subseteq \Omega^*$ be a set of MBs and $B, C \subseteq \Omega^*$. The association rule is generalized in Ω^* as followings:

The strong association and weak association of C to B in the setof MBs A is denoted by $B \rightarrow_{min} C$ and $B \rightarrow_{max} C$, respectively. An association rule of C from B is characterized by the confidence of the rule that is the ratio of the number of MBs in A exceeding every MB in both B and C to the number of MBs in A exceeding every MB in B only. More exactly, let

$$Conf_A(B \rightarrow_{min} C) = \frac{|\, supp_A^{min}(B \cup C)|}{|\, supp_A^{min}(B)|}$$

and similarly

$$Conf_A(B \rightarrow_{max} C) = \frac{|\, supp_A^{max}(B \cup C)|}{|\, supp_A^{max}(B)|}.$$

One can remark that the confidence of the association rule can be compute in polinomial time. As an consequence, we have:

Theorem 6:
There is an polinomial algorithm that for a set of MBs $A \subseteq \Omega^*$ and $B, C \subseteq \Omega^*$ decides if the association rule $B \rightarrow_{min} C$ and $B \rightarrow_{max} C$ are confident for a given threshold value $\varepsilon \geq 0$.

By previous example, let $\beta_{milk} = (1,0,0,0,0)$ and $\beta_{hammer} = (0,0,0,0,1)$ we have $supp_A^{min}(\beta_{milk}) = \frac{2}{5}$, $supp_A^{min}(\beta_{milk}, \beta_{hammer}) = 0$. Thus $Conf_A(\beta_{milk} \rightarrow_{min} \beta_{hammer}) = 0$. In words, the association of hammer to milk is quite unconfident.

Information-deficient Items in Market Baskets

The information-deficient items cause much troubles in the management. Therefore, naturally, two important problems are to determine the set of information-deficient items in a set of MBs and to determine the set of information-deficient MBs.

For $\alpha = (\alpha[1], \alpha[2], \dots, \alpha[n]) \in \Omega^*$ where Ω^* is the set of generalized MBs over the set of items $P = \{p_1, p_2, \dots, p_n\}$ let $ID(\alpha) = \{p_i | \alpha[i] = *\}$. $ID(\alpha)$ denotes the set of all items at which the manager has no information.

For a set of MBs $A \subseteq \Omega^*$ let $ID^{max}(A) = \bigcup_{\alpha \in A} ID(\alpha)$ and $ID^{min}(A) = \bigcap_{\alpha \in A} ID(\alpha)$. We denote also:

$$d^{max}(A) = \frac{|ID^{max}(A)|}{n}$$

and

$$d^{min}(A) = \frac{|ID^{min}(A)|}{n}.$$

For given threshold value $\varepsilon \geq 0$ a set of MBs A is strongly or weakly information-deficient, if $d^{max}(A) > \varepsilon$ or $d^{min}(A) > \varepsilon$, respectively.

One can remark that if $|A| = m$, $|P| = n$ then $ID^{max}(A)$ and $ID^{min}(A)$ can be computed in $O(n \times m)$ steps. Therefore in $O(n \times m)$ steps we can decide if an given set of MBs A is strongly information-deficient or weakly information-deficient. We have:

Theorem 7*:*
There is an polinomial algorithm that for a given threshold value $\varepsilon \geq 0$ and for a set of MBs $A \subseteq \Omega^*$ decides if A is strongly or weakly information-deficient.

Algorithm 1: (**Determination of the information-deficience of a given set of MBs**)

Input: A set of items $P = \{p_1, p_2, \dots, p_n\}$, a set of MBs $A = \{\alpha_1, \alpha_2, \dots, \alpha_m\} \subseteq \Omega^*$, a thershold value $\varepsilon \geq 0$.

Output: Answer to the question if A is strongly information-deficient.

Begin
Step 1: For each $\alpha_i \in A$ determine $ID(\alpha_i)$.
Step 2: Determine $ID^{max}(A) = \bigcup_{\alpha_i \in A} ID(\alpha_i)$ and $d^{max}(A) = \frac{|ID^{max}(A)|}{n}$
Step 3: If $d^{max}(A) \geq \varepsilon$, then Answer:="Yes",
else Answer:="No".
End

By the similar algorithm one can check if a set of MBs is weakly information-deficient.

On the other hand, it is important for managers to determine the information-deficient items. Let us denote:

$$D_A(p_i) = \frac{|\{\alpha \in A | \alpha[i] = *\}|}{|A|}$$

For given threshold value $\varepsilon \geq 0$ an item p_i is information-deficient in A if $D_A(p_i) > \varepsilon$.
Let us denote the set of all information-deficient items for the given threshold value $\varepsilon \geq 0$ and a set of MBs A by IDP_A^{ε}:

$$IDP_A^{\varepsilon} = \{p_i \in P | D_A(p_i) > \varepsilon\}.$$

We have also:

Theorem 8:

a) There is an polinomial algorithm that for a given threshold value $\varepsilon \geq 0$ and for a set of MBs $A \subseteq \Omega^*$ decides if an item p_i is information-deficient in A.
b) There is an polinomial algorithm that generates the set of all information-deficient items $\{p_i \in P | D_A(p_i) > \varepsilon\}$ for a given threshold value $\varepsilon \geq 0$ and for a set of MBs $A \subseteq \Omega^*$.

By a simple algorithm one can determine in polinomial time the information-deficience of a given item:

Algorithm 2:** **(**Determination of the information-deficience of an item**)

Input: A set of items $P = \{p_1, p_2, \ldots, p_n\}$, a set of MBs $A = \{\alpha_1, \alpha_2, \ldots, \alpha_m\} \subseteq \Omega^*$, a thershold value $\varepsilon \geq 0$ and an item $p_i \in P$.

Output: Answer to the question if p_i is information-deficient in A.

Begin

Step 1: Determine $D_A(p_i) = \frac{|\{\alpha \in A | \alpha[i] = *\}|}{|A|}$.

Step 2: If $D_A(p_i) \geq \varepsilon$, then Answer:="Yes",
else Answer:="No".

End

Algorithm 3:** **(**Generation of all information-deficient items**)

Input: A set of items $P = \{p_1, p_2, \ldots, p_n\}$, a set of MBs $A = \{\alpha_1, \alpha_2, \ldots, \alpha_m\} \subseteq \Omega^*$, a thershold value $\varepsilon \geq 0$.

Output: IDP_A^s.

Begin

Step 1: $IDP_A^s := \emptyset$.

Step 2: For $i = 1, \ldots, n$

Determine $D_A(p_i) = \frac{|\{\alpha \in A | \alpha[i] = *\}|}{|A|}$.

If $D_A(p_i) \geq \varepsilon$, then $IDP_A^s := IDP_A^s \cup \{p_i\}$

else i:=i+1

Output IDP_A^s.

End

Example:

Let $A = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$ be a set of MBs as in Table 1. and $B = \{\alpha_1, \alpha_2\}$. One can see that $ID^{max}(A) = \{p_3, p_4\}$, $ID^{max}(B) = \{p_3\}$ and $d^{max}(A) = \frac{2}{5}$, $d^{max}(B) = \frac{1}{2}$. Thus for $\varepsilon = 0{,}45$ A is not strongly information-deficient, while B is strongly information-deficient.

One can see that $D_A(p_1) = D_A(p_2) = D_A(p_5) = 0$, $D_A(p_3) = \frac{1}{5}$, $D_A(p_4) = \frac{2}{5}$. Thus for $\varepsilon = 0{,}3$ the set of information-deficient items is $IDP_A^s = \{p_4\}$.

Conclusion

In this research the market baskets are studied in information-deficient environment based on the previousely proposed market basket model. The concept of frequency of items and frequent items, the confidency of association rules, as well as the confident association rules are defined in the new and more generalized model. It is shown that though the frequency of items and confidency of association rules can not be envaluated exactly in an environment with insufficient information, they can be estimated approximately. The information-deficient items and information-deficient set of market baskets are also studied. It is shown that the information-deficient items and information-deficient set of market baskets can be determined. These results support the managers in management of the systems with insufficient information.

References

[1] R. Agrawal - R. Srikan (1994): Fast algorithms for mining association rules. VLDB, 487-499.

[2] T. Brüggermann - P. Hedström - M. Josefsson (2004): Data mining and Data Based Direct Marketing Activities, Book on Demand GmbH, Norderstedt, Germany.

[3] Demetrovics János, Hua, Nam Son, Guban Akos(2011) : An algebric approach to market basket model: explicit representation of frequent market basket and association rules. In: Yu Shoukourian (Ed.) Proceedings of the 8th International Conference on Computer Science and Information Technologies (CSIT 2011). pp. 170-173. ISBN:978-5-8080-0797-0.

[4] Hannu Toivonen (1996): Sampling Large Databases for Association Rules, 134--145, Morgan Kaufmann Pub.

[5] Heikki Mannila - Hannu Toivonen (1996): Discovering generalized episodes using minimal occurrences. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD' 96), AAAI Press, 146 - 151. pp.

[6] Jiawei Han - Micheline Kamber 82006): Data mining: concepts and techniques, Morgan Kaufmann Pub., (Second edition).

[7] N. Pasquier - Y. Bastide - R. Taouil - L. Lakhal (1999): Discovering frequent closed itemsets for association rules. ICDT, 398-416.

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