On Open Problem - Semantics of the Clone Items

Juraj Macko

Dept. Computer Science
Palacky University, Olomouc
17. listopadu 12, CZ-77146 Olomouc
Czech Republic
email: {juraj.macko}@upol.cz

Abstract. There was presented a list of open problems in the Formal Concept Analysis area at the conference ICFCA 2006. The problem number seven deals with the semantics of the clone items. Namely, for whom can clone items make sense and for whom can make sense the item, which can cause, that clones disappear in the collection of itemsets. In this paper we propose the semantics behind clone items with the couple of examples. Definition of the clone items is very strict and theirs use could be very limited in the real datasets. We introduce method, how to deal with items, which properties are very near to the clones. We also have a look on the items, which causes the disappearing of the clones, or decrease (increase) the degree of property "to be clone". In the experiment part we analyze some known datasets from the clone items point of view. The results bring a couple of new questions for the future research.

Keywords: formal concept analysis, clone items

1 Introduction

This paper is structured as follows: The first part, which is actually cited from the source, where the problem were defined [2] describes and defines the whole problem - the semantics of the clone items. In the second part is proposed the semantics of the clone items by putting the problem into the other point of view. There is also a discussion here, about another possible definitions of the clones as presented in [1]. In this part three comprehensive examples can be found. The third part tries to set a quite new approach to the clone items. The attributes, which are not clones, but they have properties very close to clones are considered. A nearly clones are defined. In this part some results from the introductory experiments about the clones and nearly clones are presented. Finally, the conclusion is divided in two parts - conclusion of defined problem and conclusion of other proposed issues.

2 The Problem Setting

The proposed problem of the semantics of the clone items were proposed and defined in [2] as follows: Let J be a set of items $x_1, ..., x_{|J|}$, let $\mathcal{F}$ be a collection of subsets of J and let $\varphi_{a,b}$ be the mapping $\varphi_{a,b} : 2^J \to 2^J$ defined by following formula:

$$X \to \varphi_{a,b}(X) = \begin{cases} (X \backslash \{a\}) \cup \{b\} \text{ if } b \notin X \text{ and } a \in X \\ (X \backslash \{b\}) \cup \{a\} \text{ if } a \notin X \text{ and } b \in X \\ X \text{ elsewhere} \end{cases}$$

It means swapping items a and b, which are called clone items in $\mathcal{F}$ iff for any $F \in \mathcal{F}$, we have $\varphi_{a,b}(F) \in \mathcal{F}$. A Clone-free collection is, if it does not contain any clone items.

Let (X, Y, I) be a formal context such that attributes $a \in Y$ and $b \in Y$ are not clones. Consider the formal sub-context (X, Z, I), where $Z \subset Y$, such that a and b are clone in (X, Z, I). Let $c \in Y \backslash Z$ such that a and b are no longer clone in $(X, Z \cup \{c\}, I)$. Attributes a and b has symmetrical behaviour in (X, Z, I), but this behaviour is lost when we add the attribute c to the formal context. The following question are asked:

1. Does such symmetrical behaviour of a and b make sense for someone?
2. Does it make the sense, that such symmetrical behaviour disappears, when the attribute c is added?
3. What is semantics behind the attributes a, b, and c?

3 Semantics behind Clones

3.1 Semantics behind Clones - Auxiliary Formal Definitions

The collection of itemsets will be defined as a formal context (X, Y, I), where X is a set of objects and Y is a set of attributes. Objects and attributes are related by $I \subseteq X \times Y$, which means, that the object $x \in X$ has the attribute $y \in Y$. For $A \subseteq X$, $B \subseteq Y$ and formal context (X, Y, I) we define operators

$$A^{\uparrow_I} = \{y \in Y \mid \text{ for each } x \in X \ : \ \langle x, y\rangle \in I\}$$
$$B^{\downarrow_I} = \{x \in X \mid \text{ for each } y \in Y \ : \ \langle x, y\rangle \in I\}$$

The two given attributes $a, b \in Y$ will be investigated, whether are clones or not. For this purpose the **pivot table** will be defined as the relation $R \subseteq P \times \mathcal{N}$, where $P = \{a, b\} \subseteq Y$ and $\mathcal{N}$ is a set of all N_j, where $j \in [1; |\mathcal{N}|]$. $N_j \in \mathcal{N}$ represents the set of attributes $N_j = \{x\}^{\uparrow_I} \cap (Y \backslash P)$ for each $x \in X$ such that $\{a, b\} \cap \{x\}^{\uparrow_I} \neq \emptyset$ and $\{a, b\} \nsubseteq \{x\}^{\uparrow_I}$. The investigated attributes $a, b \in P \subseteq Y$ will be named the **pivot attributes** and all other considered attributes, hence $n \in \bigcup_{j=1}^{|\mathcal{N}|} N_j$, we denote as the **non-pivot attributes**. N_j is a set generated by pivot attributes (or shortly the **generated set**). The pivot table has two rows. The "cross" × in pivot table will represent the fact, that in the formal context there exists at least one row, where the investigated attribute a (or b respectively) appears together with the attributes in the particular N_j. Formally,

$$\langle a, N_j \rangle \in R \text{ iff in context } (X, Y, I) \text{ exists } x \in X \text{ such that } x^{\uparrow_I} = \{a\} \cup N_j,$$
$$\langle b, N_j \rangle \in R \text{ iff in context } (X, Y, I) \text{ exists } x \in X \text{ such that } x^{\uparrow_I} = \{b\} \cup N_j.$$

Based on pivot attributes, non-pivot attributes and formal context (X, Y, I) consider **pivot table** which is as new formal context $(P, \mathcal{N}, R)$ with operators for $C \subseteq P$ and $\mathcal{D} \subseteq \mathcal{N}$ defined as follows

$$C^{\uparrow_R} = \{N_i \in \mathcal{N} \mid \text{ for each } p \in P : \langle p, N_j \rangle \in R\},$$
$$\mathcal{D}^{\downarrow_R} = \{p \in P \mid \text{ for each } N_i \in \mathcal{N} : \langle p, N_j \rangle \in R\}$$

In the pivot table $(P, \mathcal{N}, R)$ we are trying to find whether $\{a\}^{\uparrow_R} = \{b\}^{\uparrow_R}$. In other words, we want to know, whether the attribute a appears in given formal context with the same combination of other attributes, as b appears (in the same formal context). If yes, the pivot attributes a, b in context (X, Y, I) generates the same generated sets. In other words, a, b are not unique with respect to the non-pivot attributes. Such attributes we call clones. When $\{a\}^{\uparrow_R} \neq \{b\}^{\uparrow_R}$, attributes a, b are unique with respect to the non-pivot attributes, because generates at least one different generated set. The attribute $c \in Y$, which makes a, b unique with respect to generated sets is called the **originality factor** of a, b. In Figure 1 we show examples of the contexts and pivot tables with clones or with the originality factor respectively. By introducing the pivot table, the whole problem have been put to the other point of view. The proposed semantics will be explained based on the previous definitions.

	a	b	n_1	n_2	n_3
x_1	×		×	×	
x_2		×	×	×	
x_3	×			×	×
x_4		×		×	×
x_5	×		×		×
x_6		×	×		×
x_7	×		×	×	×
x_8		×	×	×	×

(i) Formal context (X, Z, I)

	$N_1 = \{n_1, n_2\}$	$N_2 = \{n_2, n_3\}$	$N_3 = \{n_1, n_3\}$	$N_4 = \{n_1, n_2, n_3\}$
a	×	×	×	×
b	×	×	×	×

(ii) Pivot table $(P, \mathcal{N}, R)$

	a	b	n_1	n_2	n_3	c
x_1	×		×	×		×
x_2		×	×	×		×
x_3	×			×	×	
x_4		×		×	×	
x_5	×		×		×	×
x_6		×	×		×	
x_7	×		×	×	×	
x_8		×	×	×	×	×

(iii) Formal context $(X, Z \cup \{c\}, I_c)$

	$N_1 = \{n_1, n_2, c\}$	$N_2 = \{n_2, n_3\}$	$N_3 = \{n_1, n_3\}$	$N_4 = \{n_1, n_3, c\}$	$N_5 = \{n_1, n_2, n_3\}$	$N_6 = \{n_1, n_2, n_3, c\}$
a	×	×		×	×	
b	×	×	×			×

(iv) Pivot table $(P, \mathcal{N}_c, R_c)$

Fig. 1. Formal context and pivot tables with and without clones

3.2 Discussion and Remarks

Before the comprehensive examples will be proposed, it is necessary to discuss previous auxiliary definition of the clones using the pivot table. There are couple of problems mainly dealing with ambiguity of the pivot table definition with respect to the various definitions of the clones used by the several authors in the other works. In the pivot table definition the set $N_j \in \mathcal{N}$ is defined as $N_j = \{x\}^{\uparrow_I} \cap (Y \backslash P)$ for each $x \in X$ such that

1. $\{a, b\} \cap \{x\}^{\uparrow_I} \neq \emptyset$ and
2. $\{a, b\} \nsubseteq \{x\}^{\uparrow_I}$.

The first condition tells, that we ignore the itemsets (rows), where neither a nor b is present. Such items are not interesting when we investigate whethe a and b are clones, so we will ignore them when the pivot table is defined. The second condition excludes itemsets, where we have the both pivot attributes a and b and the question is: Why we exclude such itemsets from pivot table, when we can see it in original definition of the clone items? Recall the original definition of the clones:

$$X \to \varphi_{a,b}(X) = \begin{cases} (X \backslash \{a\}) \cup \{b\} \text{ if } b \notin X \text{ and } a \in X \\ (X \backslash \{b\}) \cup \{a\} \text{ if } a \notin X \text{ and } b \in X \\ X \text{ elsewhere} \end{cases}$$

Items a and b, which are called clone items in $\mathcal{F}$ iff for any $F \in \mathcal{F}$, we have $\varphi_{a,b}(F) \in \mathcal{F}$. So we need to have the original itemset and swapped itemset as well in the whole collection of itemsets. In definition of φ are interesting the rows 1 and 2. The row 3 is only technical condition. It means, that fulfillment of swapping condition of itemsets, which does not contain any of a or b or conversely, when it contains both, is trivial. So we could add them in the pivot table by skipping the condition $\{a, b\} \nsubseteq \{x\}^{\uparrow_I}$, but we consider such information redundant and hence useless. However, the semantics of the clones remains unchanged. But on the other side, it can influence the value of the degree of clones $d^I_{(a,b)}$ (which will be defined later). In such case we need to investigate, which definition would be more precise for the user. The basic idea of our semantics of clones (and nearly clones defined later as well) is, of how original are items a and b in the whole collection of itemsets. The itemsets which does not include either a or b will not tell us anything about originality of such items, the itemsets which include both as well.

The other point for the discussion comes from the problem number six (presented in [2]), which deals with the size of a clone-free Guigues-Duquenne basis. Namely, whether the clone items are responsible for the combinatorial explosion of some Guigues-Duquennes basis. The Guigues-Duquennes basis is nonredundant. All other attribute implications, which holds in given context, can be derived from this base. In the paper [1] there are presented some partial results, which includes definitions and propositions dealing with the clones. The clones are defined with respect to pseudo-closed sets in the collection of the closed itemsets. The one of the basic results is, that in order to detect clone items, one has to consider meet-irreducible itemsets only (for details see [1]). The definition of the clone items given in [2] is defined in more general manner. It is based not only on the pseudo-closed itemset collection, but it is defined for arbitrary collection of itemsets. This fact can cause, that two items may not appear as clone according the definition in [2], but the are still clones in definition according to [1]. In the rest of the paper there will be considered the definition used in [2] only. However, the proposed semantics would be slightly modified, when we would need to use it in the meaning of [1].

The other important part is to compare proposed solution with other attempts or solutions, but the author has no information either about such attempts or about some real solutions. Hence, according to the author's best knowledge, the author's proposed solution seems to be novel.

3.3 Semantics behind Clones - Examples

In this part we would like to show on couple of examples, how the clone items and the originality factor can be used. The originality factor can be desired under some conditions, but undesired under the other conditions. Inall examples the same formal context and the pivot tables will be used, but always with the different meaning of the objects and attributes. The Table 1 represents the original formal context (X, Z, I) with the clones a and b and it also represents the formal context $(X, Z \cup \{c\})$, where the originality factor c is added. The corresponding pivot tables $(P, \mathcal{N}, R)$ and $(P, \mathcal{N}_c, R_c)$ can be seen in the Table 2. A labeling of the objects and the pivot attributes is done according to the particular sets X and Y defined in each example below.

The sales analysis Let $X = \{Customer1, \ldots, Customer8\}$ be a set of customers and the set of attributes is defined as $Y = \{Man, Woman, n_1, n_2, n_3, c\}$. The attributes Man and $Woman$ represents the sex of customer and the other attributes represents the products bought by each customer. The following formal contexts represents a marketing research of the sales company (the customers and theirs attributes). In the formal context (X, Z, I) attributes Man and $Woman$ are clones. In the pivot table $(P, \mathcal{N}, R)$ attributes Man and $Woman$ are pivot attributes and n_j is product bought by customer. On the other hand, in the formal context $(X, Z \cup \{c\})$ and the corresponding pivot table $(P, \mathcal{N}_c, R_c)$ the attributes Man and $Woman$ are no longer clones and the attribute c (Product c) is the originality factor in this case. Namely, for the itemset $\{Man, n_1, n_3, c\}$ there is no corresponding itemset $\{Woman, n_1, n_3, c\}$.

How can this information be used for the marketing department? Imagine, that the sales company wants to create packages based on the marketing research. These packages should consist of the particular products n_j. In the first

	Man / *Europe* / *Gene*1	*Woman* / *America* / *Gene*2	n_1	n_2	n_3	c
Customer 1 / *Animal* 1 / *Organism* 1	×		×	×		×
Customer 2 / *Animal* 2 / *Organism* 2		×	×	×		×
Customer 3 / *Animal* 3 / *Organism* 3	×			×	×	
Customer 4 / *Animal* 4 / *Organism* 4		×		×	×	
Customer 5 / *Animal* 5 / *Organism* 5	×		×		×	×
Customer 6 / *Animal* 6 / *Organism* 6		×	×		×	
Customer 7 / *Animal* 7 / *Organism* 7	×		×	×	×	
Customer 8 / *Animal* 8 / *Organism* 8		×	×	×	×	×

Table 1. Formal contexts (X, Z, I) and $(X, Z \cup \{c\}, I_c)$

	$\mathcal{N}$				$\mathcal{N}_c$					
	$N_1 = \{n_1, n_2\}$	$N_2 = \{n_2, n_3\}$	$N_3 = \{n_1, n_3\}$	$N_4 = \{n_1, n_2, n_3\}$	$N_1 = \{n_1, n_2, c\}$	$N_2 = \{n_2, n_3\}$	$N_3 = \{n_1, n_3\}$	$N_4 = \{n_1, n_3, c\}$	$N_5 = \{n_1, n_2, n_3\}$	$N_6 = \{n_1, n_2, n_3, c\}$
Man / *Europe* / *Gene*1	×	×	×	×	×	×		×	×	
Woman / *America* / *Gene*2	×	×	×	×	×	×	×			×

Table 2. Pivot tables $(P, \mathcal{N}, R)$ and $(P, \mathcal{N}_c, R_c)$

case of the formal context (X, Z, I) the company can create the same packages for man and for woman, because male and female customers buy the same combinations of products n_j. The same packages for two different groups can reduce the total cost of production, because we need to produce only four types of the packages, namely the packages $N_1 = \{n_1, n_2\}$, $N_2 = \{n_2, n_3\}$, $N_3 = \{n_1, n_3\}$ and $N_4 = \{n_1, n_2, n_3\}$. With the attribute c added to the formal context, we need six different packages, because only the packages $N_1 = \{n_1, n_2\}$ and $N_2 = \{n_2, n_3\}$ can be produced for men and women at the same time. Other packages are different for the male and female customers. From this point of view, the originality factor is undesired and the clones are desired.

But we can use this information in the other way. Suppose, that the cost difference of producing four or six package types is not significant, but significant can be a targeted marketing on the male and female customers. The formal context (X, Z, I), where we have the clone attributes Man and $Woman$, does not provide differentiated information about the male and female customers. On the other hand, the formal context $(X, Z \cup \{c\})$ does. The attribute c provides desired information, that the Product c influences the different combination of the products bought by the male and female customer. It means, that we can make targeted marketing (namely, the different type of packages for the different type of customers) based on the originality factor Product c and its combinations with the other products. Some combinations of the products with the originality factor can be used as a topic for advertising to highlight the difference between man and woman preferences. However, the clone analysis can provide the marketing department with the useful information in both cases.

Analysis of the animals Let $X = \{Animal1, \ldots, Animal8\}$ be a set of animals and a set of attributes is defined as $Y = \{Europe, America, n_1, n_2, n_3, c\}$. The formal context (X, Z, I) in the Table 1, shows the attributes $Europe$ and $America$ as clones. This fact can be interpreted as follows: In Europe and in America they live the same types of animals, when we consider the attributes of the animals n_1, n_2 and n_3 only. The same information can be seen in the pivot table Table 2. When we add the attribute c, we can see the different types of animals (with the different generated sets) in Europe and in America as well (see Table 2). The information, that exists the originality factor c for attributes $Europe$ and $America$ can be interpreted as follows: It shows, that $Europe$ and $America$ are somehow specific. In Europe are some different combinations of animal's attributes than in America and vice versa and at the same time we see, that this difference somehow deals with the attribute c. Biologist can investigate in more details, what is specific in Europe and in America, which specific attribute of Europe leads to the different attributes of the animals in Europe (and vice versa). Other use of such information is following: From a background knowledge we know, that there is no reason for differentiating the animals in Europe and America just on attribute c. In our dataset we do not have in America the animal with attributes n_1, n_3 and c, but with respect to the attribute c we expect to have the same types of animals in Europe and in America. Thus, we need to look for such animal in America as well. Our hypothesis is, that in America lives such animal, because it lives in Europe and based on our background knowledge there is no reason for c to be the originality factor. From the formal point of view, we do not have the complete dataset (formal context). Some rows are missing, and we need to find such objects in the reality (in this case we are looking for the animal).

Analysis of genes and the morphological attributes of organisms The last example use set $X = \{Organism1, \ldots, Organism8\}$ and set of attributes $Y = \{Gene1, Gene2, n_1, n_2, n_3, c\}$ The attributes $Gene1$ and $Gene2$ are clones in formal context (X, Z, I) in Table 1 and the other attributes represents the morphological property of the organism. The interpretation can be following: Organisms with $Gene1$ and $Gene2$ has the same combination of morphological properties N_j, when we consider the morphological properties of organisms n_1, n_2 and n_3. The same information can be seen in the pivot table Table 2. When we add the morphological attribute c, we get the formal context $(X, Z \cup \{c\})$, which means, that based on attribute c there are some different types of the morphological attributes of organisms with the $Gene1$ and $Gene2$ (see the Table 1 and Table 2). It shows, that the $Gene1$ and $Gene2$ probably does not influence the sets of the morphological attributes containing only n_1, n_2, n_3, but this $Gene1$ and $Gene2$ influence the sets of the morphological attributes containing c. Thus, c as the originality factor makes the difference between these two genes. This information could be useful for a hypothesis creation in genetics.

4 Nearly Clones

4.1 Degree of Clones and Degree of Originality

The definition of the clone items is very strict. Recall, that condition $\varphi_{a,b}(F) \in \mathcal{F}$ needs to be true for any $F \in \mathcal{F}$. We can see, that adding only one "cross" into the huge formal context can cause, that two clones disappear. We expect, that in real dataset such condition can be true very rarely. When we want to use the clone items meaningfully, we need to have a weaker definition. For practical purposes it suffices, that condition $\varphi_{a,b}(F) \in \mathcal{F}$ can be true in some reasonable amount of $F \in \mathcal{F}$. We define **degree of clone** as

$$d^I_{(a,b)} = \frac{|\{a\}^{\uparrow_R} \cap \{b\}^{\uparrow_R}|}{|\{a\}^{\uparrow_R} \cup \{b\}^{\uparrow_R}|},$$

which can be read as follows: The attributes a and b with respect to the formal context I are clones in the degree d. For a priori given threshold θ we define a and b as **nearly clones** iff $d_{(a,b)} \geq \theta$. Note, that for $d_{(a,b)} = 1$ the attributes a and b are clones and for $d_{(a,b)} = 0$ we say, that they are **original attributes**. Consider now the formal context (X, Z, I_Z) and the corresponding pivot table $(P, \mathcal{N}_Z, R_Z)$ (see Figure 2). We can see, that a and b are clones with the degree

$d^{I_Z}_{(a,b)} = 1$. Adding either attribute c_1 or attribute c_2 to the formal context leads to decreasing of clone degree for a and b. Namely, $d^{I_{c_1}}_{(a,b)} = 0$ and $d^{I_{c_2}}_{(a,b)} = 0,6$. In both cases degree has decreased, but the resulted clone degree is different. In the first case attributes are original, in the second case attributes are nearly clones for arbitrary $\theta \leq 0,6$. Such situation can be formalized, and define the **degree of originality** for given c_i and context (X, Z, I) as

$$g^{I_c}_{(a,b)} = d^{I}_{(a,b)} - d^{I_c}_{(a,b)}$$

The degree of originality shows, how the attribute, added to the context, does influence the degree of clone for given attributes $a, b \in Y$ and the formal context (X, Z, I).

	a	b	n_1	n_2	n_3	c_1	c_2
x_1	×		×	×		×	×
x_2		×	×	×			
x_3	×			×	×	×	
x_4		×		×	×		
x_5	×		×		×	×	
x_6		×	×		×		
x_7	×		×	×	×	×	
x_8		×	×	×	×		

(i) Formal contexts (X, Z, I_Z), $(X, Z \cup \{c_1\}, I_{c_1})$, and $(X, Z \cup \{c_2\}, I_{c_2})$

	$N_1 = \{n_1, n_2\}$	$N_2 = \{n_2, n_3\}$	$N_3 = \{n_1, n_3\}$	$N_4 = \{n_1, n_2, n_3\}$
a	×	×	×	×
b	×	×	×	×

(ii) Pivot table $(P, \mathcal{N}_Z, R_Z)$ from context (X, Z, I_Z) $d^{I_Z}_{(a,b)} = 1$

	$N_1 = \{n_1, n_2, c_1\}$	$N_2 = \{n_1, n_2\}$	$N_3 = \{n_2, n_3, c_1\}$	$N_4 = \{n_2, n_3\}$	$N_5 = \{n_1, n_3, c_1\}$	$N_6 = \{n_1, n_3\}$	$N_7 = \{n_1, n_2, n_3\}$	$N_8 = \{n_1, n_2, n_3, c_1\}$
a	×		×		×		×	
b		×		×		×		×

(iii) Pivot table $(P, \mathcal{N}_{c_1}, R_{c_1})$ from context $(X, Z \cup \{c_1\}, I_{c_1})$ $d^{I_{c_1}}_{(a,b)} = 0$

	$N_1 = \{n_1, n_2, c_2\}$	$N_2 = \{n_1, n_2\}$	$N_3 = \{n_2, n_3\}$	$N_4 = \{n_1, n_3\}$	$N_5 = \{n_1, n_2, n_3\}$
a	×		×	×	×
b		×	×	×	×

(iv) Pivot table $(P, \mathcal{N}_{c_2}, R_{c_2})$ from context $(X, Z \cup \{c_2\}, I_{c_2})$ $d^{I_{c_2}}_{(a,b)} = 0,6$

Fig. 2. Formal contexts and pivot tables with different degrees of clone.

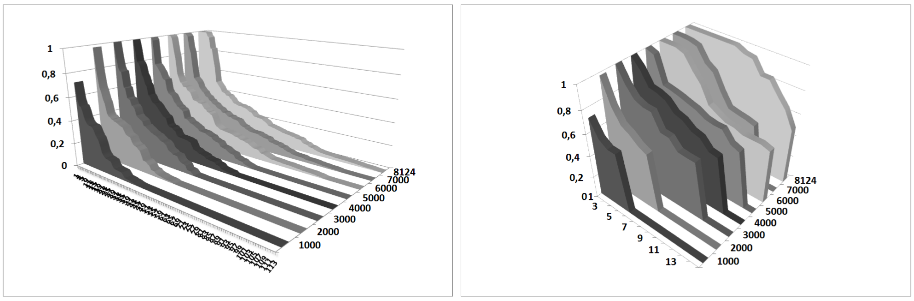
4.2 Experiment Nr. 1 - Amounts and Degrees of Nearly Clones in Datasets

For the purpose of this paper we arranged two introductory experiments with the nearly clones, in which we use datasets *Mushroom* [3], *Adults* [4] and *Anonymous* [5] from well known UC Irvine Machine Learning Repository (for the details see Table 3). In the experiments we used a naive algorithm (the brute-force search, but with polynomial complexity) for finding the degrees of clones as defined above. Looking for more efficient algorithm is out of scope of this paper. The algorithm was implemented in C, and all experiments have been run on the computer with an Intel Core i5 CPU, 2.54 Ghz, 6 GB RAM, 64bit W7 Professional.

In the first experiment we were focused on finding all nearly clone pairs, with $d_{(a,b)} > 0$, especially we investigated, if there are some clones (where $d_{(a,b)} = 1$) in the real datasets. The results of the first experiment are shown in Table 3.

	Mushroom [3]	Adults [4]	Anonymous [5]
Number of objects	8 124	48 842	32 713
Number of attributes	119	104	295
Number of nearly clones $d_{(a,b)} > 0$	113	1 568	382
Maximal $d_{(a,b)} > 0$	1,00000	0,02252	0,00187
Minimal $d_{(a,b)} > 0$	0,00123	0,00014	0,00143
Average $d_{(a,b)} > 0$	0,24423	0,00449	0,00160
Median $d_{(a,b)} > 0$	0,14537	0,00195	0,00159
Slope	0,99402	0,77895	0,55004

Table 3. Overview of the datasets and results of the first experiment (source of datasets: *http://archive.ics.uci.edu/ml/index.html*)



(i) nearly clone pairs for $d_{(a,b)} > \theta = 0$
x-axis - number of nearly clone pairs
y-axis - degree of clone $d_{(a,b)}$
z-axis - number of objects processed

(ii) nearly clone pairs for $d_{(a,b)} \geq \theta = 0.5$
x-axis - number of nearly clone pairs
y-axis - degree of clone $d_{(a,b)}$
z-axis - number of objects processed

Fig. 3. Mushroom - distribution of $d_{(a,b)}$ in dataset scaled by 1000 objects.

In case of the dataset *Mushroom*, we present also the distribution of the clone degrees and some other details as well. Figure 3 shows the volume of all pairs a and b and clone degree $d_{(a,b)} > 0$, for each scale pattern (from 1000 to 8124 by 1000). In (i) are displayed all pairs with $d_{(a,b)} > 0$ and part (ii) is more focused on the amount of pairs where $d_{(a,b)} \geq 0,5$ for each investigated scaled pattern. Note, that the results from numbers of the processed objects in the dataset *Mushroom* (namely from 1000 to 7000 depicted in z-axis in the Figure 3) depends on an order of the processing objects. This fact were not investigated more deeply. However, when we have processed all 8124 objects, the order will not influence the result. Figure 4 shows some interesting details. In (i) there are presented the pairs a, b with $d_{(a,b)} = 1$, in (ii) the same for $1 > d_{(a,b)} \geq 0,5$. We have found 4 clone pairs, and one clone triple. In the clone triple $(103, 104, 105)$ we can see the transitivity (i.e. when (a, b) are clones and b, c are clones, also a and c are clones). Such transitivity is not surprising and is direct consequence of the clone definition.

What does such results show and does it appear reasonable? The Figure 4 part (i) shows the clones a and b. The original dataset $Mushroom$ consists of 22 attributes with non-binary values. For the purposes of clone investigation, this dataset were nominally scaled to the formal context, which is binary indeed. It is interesting to see, that all clone items represents the value of the same original attribute. E.g. clones 019 and 021 represents the original attribute *Cap Color*, thus its values *Purple* or $White$ respectively. Another example is clone triple 103, 104 and 105 which represents the original attribute *Spore Print Color* with the corresponding values *Orange*, *Purple* and $White$. It can be interpreted as follows: Purple and white color generates the same sets of the non-pivot attributes. In other words, to each mushroom with the purple cap (the pivot attribute), there exists corresponding mushroom with the white cap (the pivot attribute), but all other properties remains he same (non-pivot attributes). Similarly to each mushroom with the purple spore print color, there exists corresponding mushroom with the white spore print color and the corresponding mushroom with the orange one. When we look on the nearly clones in the Figure 4 part (ii), the attributes 69 and 70 represents the same original attribute *stalk color above ring* with the values *cinnamon* and *gray* (the details are not shown in the table). These attributes are not the clones, but the nearly clones with the clone degree $d_{(a,b)} = 0,96$. It can be interpreted similarly as by the clones. Only the difference will be in a quantifier. By clones the quantifier was "for each", by the nearly clones we will have fuzzy quantifier, in this case "for the most". Hence the interpretation is: For the most mushroom with cinnamon stalk above the ring exists corresponding mushroom with the corresponding gray stalk above the ring (and vice versa). The clone degree is very high in this case ($d_{(a,b)} = 0,96$) it means there are only couple of mushrooms with cinnamon stalk above the ring color, which do not have corresponding mushroom with the gray stalk above the ring color. However, the for the deeper understanding of such examples, it is required to ask an expert in mycology.

4.3 Experiment 2 - Structure of Nearly Clones in Datasets

In the second experiment we investigated the structure of nearly clones. Namely, we have defined a fuzzy relation $T : Y \times Y \to L$, where $L = [0; 1]$ is defined as $T(a, b) = d_{(a,b)} \in L$. In other words, the fuzzy relation express the degree of the clone for each pair $a, b \in Y$. For the better visualization we display such relation in so called "bubble chart". The bubble chart displays three dimensional data in two dimensional chart. The position of the bubble is given by two dimensions (x and y axis) and the size of the bubble shows the third dimension. The results from the first experiments are displayed in bubble chart, where the pairs of attributes a and b represents two dimensions and the degree of clone $d_{(a,b)}$ is represented by the size of the bubble. Note, that the fuzzy relation T is indeed symmetric (i.e. $d_{(a,b)} = d_{(b,a)}$), but we show only part of the relation, where $a < b$. Figures 5, 6 and 7 show the structure of the nearly clones for the datasets *Mushroom*, *Anonymous* and *Adults*. We can observe very different structure of the nearly clones in each dataset. In *Mushroom* we can see, that the structure of the nearly clones is approximately linear. All nearly clones are clustered near to the line defined as (y, y). In the case of *Adults* dataset we can see more spread, but still approximately linear structure, except of one cluster near point $(0, |Y|)$. The nearly clones of the data set *Anonymous* forms the different, but kind of regular structure as well. This results leads to the question, what kind of properties has fuzzy relation of the nearly clones and if properties of such fuzzy relation correlates with the properties of the formal context, or with properties of the concept lattice. Until now we know, that such relation is transitive for clone items $d_{(a,b)} = 1$ and symmetric for the arbitrary nearly clone items, but this two properties are trivial. I would be also interesting to find the semantics of such fuzzy relation defined on the nearly clones. All this will be part of the future investigation.

original attributes	a	b	$d_{(a,b)} = 1$
03. cap-color	019 purple=u	021 white=w	1,00
05. odor	025 almond=a	026 anise=l	1,00
05. odor	030 musty=m	031 none=n	1,00
17. veil-color	086 brown=n	087 orange=o	1,00
20. spore-print-color	103 orange=o	104 purple=u	1,00
20. spore-print-color	103 orange=o	105 white=w	1,00
20. spore-print-color	104 purple=u	105 white=w	1,00

(i) Clones

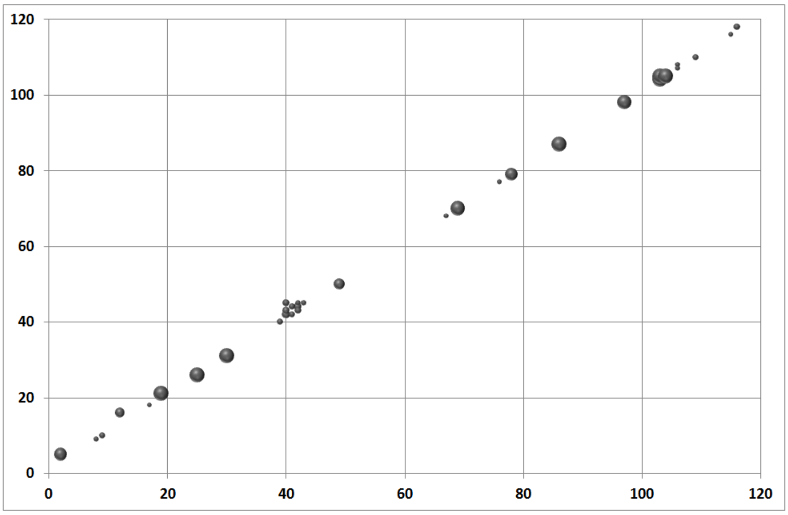
a	b	$d_{(a,b)} \geq 0.5$
69	70	0,96
97	98	0,95
78	79	0,84
2	5	0,84
49	50	0,74
12	16	0,66
40	42	0,51

(ii) Nearly clones

Fig. 4. Dataset Mushroom - experiment on nearly clones

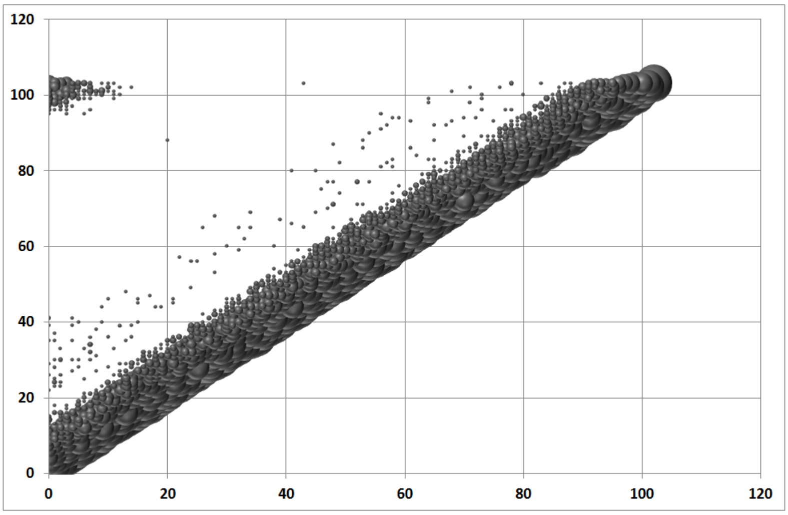
5 Conclusion and Future Perspectives

The paper was motivated by open problem proposed at ICFCA 2006 [2]. We hope, that this small open problem is solved now and the reason is presented in the first part of the conclusion. This part is structured as a direct answers on proposed questions. The second part of the conclusion describes ideas, which overlaps the original open problem and come with some new questions.



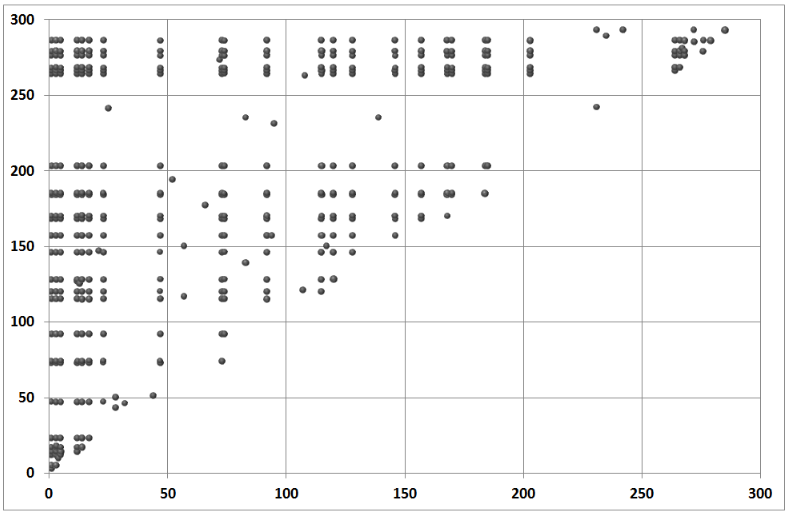
Nearly clone pairs for $d_{(a,b)} > \theta = 0$, where x,y-axis - a and b pairs size of bubble $= d_{(a,b)}$

Fig. 5. Dataset Mushroom



Nearly clone pairs for $d_{(a,b)} > \theta = 0$, where x,y-axis - a and b pairs size of bubble $= d_{(a,b)}$

Fig. 6. Dataset Adults



Nearly clone pairs for $d_{(a,b)} > \theta = 0$, where x,y-axis - a and b pairs
size of bubble = $d_{(a,b)}$

Fig. 7. Dataset Anonymous

5.1 Conclusion for Open Problem Questions

Question 1: Does the symmetrical behaviour of a and b make sense for someone?
Answer 1: Yes, such symmetrical behaviour can identify the same combination of the non-pivot attributes with respect to pivot attributes and can make sense:

- for the marketing department to reduce cost of packages - the clone items enable the same packages for the different types of customers (e.g. man and woman)
- for biologists to complete the the dataset - the clone items are expected, because the originality factor c, has no sense based on the background knowledge. Hence, we some miss rows in the dataset (e.g. we need to find the new animals)
- for genetics - it bring an information that the two genes has no influence on a combination of the morphological properties of organisms
- generally for everyone, who needs an information about the same combination of non-pivot attributes with respect to the pivot attributes

Question 2: Does it make sense, that such symmetrical behaviour disappear, when c is added?
Answer 2: Yes, such attribute is called the originality factor for the items a and b and can be useful:

- for the marketing department to make a targeted marketing for the different types of customers (e.g. man and woman) using unique combination of the non-pivot attributes
- for biologist to find the difference between two pivot attributes (e.g. Europe and America) with respect to other non-pivot properties. The originality factor c reveals, that the pivot attributes are original and this originality needs to be investigated deeper.
- for genetics - it brings an information, that two genes has an influence on a combination of the morphological properties of organisms
- generally for everyone, who needs an information about the reason, why the non-pivot attributes has the different combinations with respect to the pivot attributes.

Question 3: What is semantics behind a, b, and c?
Answer 3: The attributes a and b are the pivot attributes, all other attributes are the non-pivot attributes and c is moreover the originality factor for the attributes a and b. The pivot attributes generates a combination of the non-pivot attributes in the given context. The attribute c make the attributes a and b unique, which can be "good" or "bad". It depends on a goal of the analysis.

5.2 Conclusion and Future Perspectives

The second part of conclusion shows, that the clones are very strictly defined. Therefore the nearly clones were introduced. The nearly clones operates with the degree, in which two attributes are clones. Such formalization asks itself for study of the nearly clones under fuzzy setting (e.g. we have already mentioned, that structure of nearly clones can be seen as fuzzy relation indeed). The introductory experiments shows, that the nearly clones in dataset have an interesting structure, which needs to be investigated more deeply. This paper was introductory for the nearly clones. As a future work we plan to describe more efficient algorithm to compute the nearly clones for the given threshold θ, and algorithm for identifying the originality factors for another given threshold ω. Finally we hope, that this paper, even it does not come with a great mathematical or experimental results, brings some interesting ideas to FCA community.

Acknowledgements. The author is very grateful to the reviewers for their helpful comments and suggestions. Partly supported by IGA (Internal Grant Agency) of the Palacky University, Olomouc is acknowledged.

References

1. Gély A,, Medina A., Nourine L. and Renaud Y.: *Uncovering and Reducing Hidden Combinatorics in Guigues-Duquenne Bases.* Springer, Lecture Notes in Computer Science, 2005, Volume 3403/2005, 235-248, Heidelberg 2005
2. more authors: *Some open problems in Formal Concept Analysis.* ICFCA 2006, Dresden, http://www.upriss.org.uk/fca/fcaopenproblems.html

3. Schlimmer,J.S. : *Concept Acquisition Through Representational*, Adjustment (Technical Report 87-19). (1987). Doctoral disseration, Department of Information and Computer Science, University of California, Irvine.
4. Kohavi R.: *Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid* Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996
5. Breese J., Heckerman D., Kadie C.: *Empirical Analysis of Predictive Algorithms for Collaborative Filtering.* Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, July, 1998.