On Stability of Triadic Concepts

Sergei O. Kuznetsov[1] and Tatiana P. Makhalova[1,2]

[1]National Research University Higher School of Economics, Kochnovsky pr. 3,
Moscow 125319, Russia
[2]ISIMA, Complexe scientifique des Cézeaux, 63177 Aubière Cedex, France

skuznetsov@hse.ru,t.makhalova@gmail.com

Abstract. Triadic formal concept analysis has become a popular research direction, since triadic relations give natural models of many data collections. In this paper we address the problem of selecting most interesting concepts by proposing triadic stability indices.

1 Introduction

Triadic formal concept analysis (3FCA) was introduced by Rudolf Wille and Fritz Lehmann [1] to model hierarchies of classes and dependencies arising from ternary relations. Recently, several algorithms for computing frequent triconcepts were proposed [2, 3]. It is noticed that some infrequent concepts are still interesting, since they represent extraordinary or uncommon data. In this paper we propose triadic stability for selecting interesting triadic concepts. Together with exact stability indices we suggest their efficient approximations analogous to Δ-stability introduced in [4].

2 Main definitions

2.1 Formal Concept Analysis

First, we briefly recall some basic definitions of the Formal Concept Analysis (FCA) [5]. A formal context is a triple (G, M, I). G and M are sets of objects and attributes respectively, and I is an incidence relation. It is defined as the Cartesian product $G \times M$, i.e. $(g, m) \in I$ if the object $g \in G$ has the attribute $m \in M$. The derivation operators $(\cdot)'$ are defined for $A \subseteq G$ and $B \subseteq M$ as follows:

$$A' = \{m \in M \mid \forall g \in A : gIm\}$$
$$B' = \{g \in G \mid \forall m \in B : gIm\}$$

A' is the set of attributes common to all objects of A, and B' is the set of objects sharing all attributes of B. The double application of $(\cdot)'$ is a closure operator, i.e. $(\cdot)''$ is extensive, idempotent and monotone. Subsets $A \subseteq G$, $B \subseteq M$ such that $A = A''$ and $B = B''$ are called *closed*.

A (formal) concept is a pair (A, B), where $A \subseteq G$, $B \subseteq M$ and $A' = B$, $B' = A$. A is called the (formal) extent, and B is called the (formal) intent of the concept (A, B).

A concept lattice (or Galois lattice) is a partial ordered set of concepts, the order $\leqslant$ on the set of concepts is defined as follows: $(A, B) \leq (C, D)$ iff $A \subseteq C\,(D \subseteq B)$, a pair (A, B) is a subconcept of (C, D), while (C, D) is a superconcept of (A, B). Each finite lattice has the highest element with $A = G$, called the top element, and the lowest element with $B = M$, called the bottom element.

2.2 Triadic Concept Analysis

In the case of a triadic relation one deals with a quadruple (G, M, B, Y), called a triadic context. G, M, B are sets and Y is a ternary relation between G, M and B, i.e. $Y \subseteq G \times M \times B$; the elements of G, M and B are called objects, attributes and conditions respectively, and $(g, m, b) \in Y$ is read: object g has attribute m under condition b.

The dyadic derivation operators can be used to construct triadic concepts. A triadic context can be represented as follows: $\mathbb{K} := (K_1, K_2, K_3, Y)$, where K_1 is a set of objects G, K_2 is a set of attributes and K_3 is a set of conditions, and each element of K_i may be seen as an instance of Peirce's i-th category [1]. For every triadic context one defines the following dyadic contexts:

$\mathbb{K}^1 := \left(K_1, K_2 \times K_3, Y^{(1)}\right)$ with $gY^{(1)}(m, b) :\Leftrightarrow (g, m, b) \in Y$

$\mathbb{K}^2 := \left(K_2, K_1 \times K_3, Y^{(2)}\right)$ with $mY^{(2)}(g, b) :\Leftrightarrow (g, m, b) \in Y$

$\mathbb{K}^3 := \left(K_3, K_1 \times K_2, Y^{(3)}\right)$ with $bY^{(3)}(g, m) :\Leftrightarrow (g, m, b) \in Y$

For $\{i, j, k\} = \{1, 2, 3\}$ and $A_k \subseteq K_k$, one defines $\mathbb{K}^{(i,j)}_{A_k} := \left(K_i, K_j, Y^{(i,j)}_{A_k}\right)$, where $(a_i, a_j) \in Y^{(i,j)}_{A_k}$ if and only if $(a_i, a_j, a_k) \in Y$ for all $a_k \in A_k$.

Put differently, the context $\mathbb{K}^{(i)}$ is a flattened representation of the original triadic context, while $\mathbb{K}^{(i,j)}_{A_k}$ corresponds to the relation between elements of K_i and K_j that belong to A_k.

(i)-derivation operator For $\{i, j, k\} = \{1, 2, 3\}$ with $j < k$ and for $X \subseteq K_i$ and $Z \subseteq K_j \times K_k$ the (i)-derivation operators are defined by

$$X \longmapsto X^{(i)} := \{(a_j, a_k) \in K_j \times K_k \mid (a_i, a_j, a_k) \in Y \text{ for all } a_i \in X\}$$

$$Z \longmapsto Z^{(i)} := \{a_i \in K_i \mid (a_i, a_j, a_k) \in Y \text{ for all}\,(a_j, a_k) \in Z\}$$

(i, j, X_k)-derivation operators For $\{i, j, k\} = \{1, 2, 3\}$ and for $X_i \subseteq K_i$, $X_j \subseteq K_j$ and $A_k \subseteq K_k$ the (i, j, X_k)-derivation operators are defined by

$$X_i \longmapsto X_i^{(i,j,A_k)} := \{a_j \in K_j \mid (a_i, a_j, a_k) \in Y \text{ for all}\,(a_i, a_k) \in X_i \times A_k\}$$

$$X_j \longmapsto X_j^{(i,j,A_k)} := \{a_i \in K_i \mid (a_i, a_j, a_k) \in Y \text{ for all}\,(a_j, a_k) \in X_j \times A_k\}$$

A triadic concept (triconcept) of $\mathbb{K} := (K_1, K_2, K_3, Y)$ is a triple (A_1, A_2, A_3) with $A_i \subseteq K_i$ for $i \in \{1, 2, 3\}$ and $A_i = (A_j \times A_k)^{(i)}$ for every $\{i, j, k\} = \{1, 2, 3\}$ with $j < k$. The sets A_1,A_2, and A_3 are called extent, intent and modus of the triadic concept respectively. We let $\mathfrak{T}(K)$ denote the set of all triadic concepts of $\mathbb{K}$.

A triadic concept lattice has three maximal elements, namely $((K_2 \times K_3)^{(1)}, K_2, K_3)$, $(K_1, (K_1 \times K_3)^{(2)}, K_3)$, and $(K_1, K_2, (K_1 \times K_2)^{(3)})$. For any two elements of a lattice one defines tree types of set inclusion/exclusion relations, which satisfy the following antiordinal dependencies: $(A_1, A_2, A_3) \preceq_G (B_1, B_2, B_3)$ iff $A_1 \subseteq B_1, A_2 \supseteq B_2, A_3 \supseteq B_3$, $(A_1, A_2, A_3) \preceq_M (B_1, B_2, B_3)$ iff $A_1 \supseteq B_1, A_2 \subseteq B_2, A_3 \supseteq B_3$ or $(A_1, A_2, A_3) \preceq_C (B_1, B_2, B_3)$ iff $A_1 \supseteq B_1, A_2 \supseteq B_2, A_3 \subseteq B_3$.

3 Stability Indices For Triadic Concepts

Stability indices for formal concepts were introduced in [6, 7] and modified in [8]. We define stability indices for the triadic case in a similar way. We describe two types of stability that correspond to the derivation operators defined above.

(i)-stability For a triadic concept (A_1, A_2, A_3) the (i)-stability is defined by:

$$Stab^{(i)}(A_1, A_2, A_3) := \frac{|\left\{X \subseteq A_i | X^{(i)} = (A_j \times A_k)\right\}|}{2^{|A_i|}}$$

This index shows how much the binary relation on sets X_j and X_k is dependent on particular elements of a subset A_i.

(i, j, X_k)-stability For a triadic concept (A_1, A_2, A_3) the (i, j, X_k)-stability is defined by:

$$Stab^{(i,j,X_k)}(A_1, A_2, A_3) := \frac{|\left\{X \subseteq (A_i \times A_j)\, | X^{(k)} = A_k\right\}|}{2^{|A_i| + |A_j|}}$$

The (i, j, X_k)-stability allows us to estimate the dependence of a subset A_k on elements of the (X_i, X_j)-relation.

Example Below we consider a small examples of computing stability indices for a concept. The formal context is given in the table 1.

Table 1. Triadic context

	α				β				γ				δ			
	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d
1	×				×			×				×				
2	×				×	×	×			×	×			×	×	
3			×	×	×	×	×			×	×					
4								×								

Let us consider a triconcept $C = (\{2,3\}, \{b,c\}, \{\beta,\gamma\})$ with (1) - stability and $(1,3,X_2)$-stability.

$Stab^{(1)}(C) = \frac{1}{2}$. Since the numerator is comprised by $\{3\}^{(1)} = (\{b,c\} \times \{\beta,\gamma\})$ and $\{2,3\}^{(1)} = (\{b,c\} \times \{\beta,\gamma\})$.

Table 2. $I \subseteq M \times C$ corresponding to all possible subsets of the extent $\{2,3\}$

$\{\emptyset\}$	a	b	c	d	$\{2\}$	a	b	c	d	$\{3\}$	a	b	c	d	$\{2,3\}$	a	b	c	d
α	×	×	×	×	α	×				α			×	×	α				
β	×	×	×	×	β	×	×	×		β	×	×	×		β	×	×	×	
γ	×	×	×	×	γ		×	×		γ		×	×		γ		×	×	
δ	×	×	×	×	δ		×	×		δ					δ				

$Stab^{(1,3,X_2)}(C) = \frac{3}{8}$. To compute this value one needs to check 16 subsets of $X_1 \times X_3$ and corresponding subsets of X_2. The following sets occur in the numerator: $\{\emptyset, 2, 3, 23\} \times \{\emptyset, \beta, \gamma, \beta\gamma\}$.

$$\{b,c\} = (\{2,3\}, \{\beta,\gamma\})^{(1,3,A_2)} = (\{2,3\}, \{\gamma\})^{(1,3,A_2)} = (\{3\}, \{\gamma\})^{(1,3,A_2)}$$
$$(\{3\}, \{\beta,\gamma\})^{(1,3,A_2)} = (\{2\}, \{\gamma\})^{(1,3,A_2)} = (\{2\}, \{\beta,\gamma\})^{(1,3,A_2)}$$

4 Estimates of stability

The problem of computing stability is $\#P$-complete [6, 7], therefore, in practice, when one deals with a big context and with the huge amount of generated concepts, it is very difficult to apply these indices. That's why, estimates of the stability for dyadic concepts have been proposed [9, 4].

We have expanded the Δ-stability [4] for the case of triadic stability indices. In this regard, it is important to note that the estimates derived from the direct descendants of a triconcept can be useless owing to the defined quasiorders, because the number of direct neighbors is usually small. In figure 1 the distributions of the descendants number with respect to different inclusion/exclusion relations are represented.

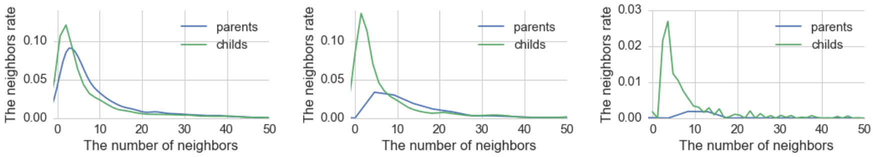

Fig. 1. The number of neighbors distributions

Instead of considering the set difference between (i)-th components of a triconcept and each direct descendant, we consider the set difference between (i)-th components of a triconcept $c = (A_1, A_2, A_3)$ and other, possibly, unclosed concepts derived by adding new elements from $K_j \setminus A_j$, $j \neq i$ or $K_j \times K_k \setminus A_j \times A_k$, $j, k \neq i$. Put differently, the lower and upper bounds estimates of stability index look as follows:

$$-log_2 \sum_{d \in Enl(c)} 2^{-\Delta(c,d)} \leq LStab\left(c\right) \leq \Delta_{min}\left(c\right),$$

where $\Delta_{min}\left(c\right) = min_{d \in Enl(c)} \Delta\left(c, d\right)$,

$$Enl\left(c\right) = \left\{X \mid X = \left\{A_k \cup x\right\}, x \in K_k \setminus A_k, X^{(k)} \subseteq \left(A_i \times A_j\right)\right\}$$

and $\Delta\left(c, d\right)$ is the difference between $|A_j| \cdot |A_k|$ and the number of elements in $X^{(k)}$ for estimates of (i, j, X_k)-stability.

$$Enl\left(c\right) = \left\{X \mid X = \left\{A_j \times A_k \cup x\right\}, x \in K_j \times K_k \setminus A_j \times A_k, X^{(i)} \subseteq A_i\right\}$$

and $\Delta\left(c, d\right)$ is the difference between $|A_i|$ and the number of elements in $X^{(i)}$ for estimates of (i)-stability.

Example. Consider upper and lower bounds of stability estimates for $C = \left(\left\{2, 3\right\}, \left\{b, c\right\}, \left\{\beta, \gamma\right\}\right)$ from the running example (Table 1). To get an estimate of the (1)-stability we consider elements of the following set $\left\{a, b, c, d\right\} \times \left\{\alpha, \beta, \gamma, \delta\right\} \setminus \left\{b, c\right\} \times \left\{\beta, \gamma\right\}$. Subsets of A_1 derived from those elements are $\left\{\emptyset\right\}$ and $\left\{2\right\}$, which give us $-log_2(7 \cdot 2^{-2} + 2^{-1})$ and 1 for lower and upper bounds, respectively. To get estimates of $(1, 3, X_2)$-stability one needs to expand the intent by elements from $\left\{a, d\right\}$. Adding the first element a reduces the $\left\{2, 3\right\} \times \left\{\beta, \gamma\right\}$ to $\left\{2, 3\right\} \times \left\{\beta\right\}$, while expanding the intent by d results in the empty set. Thus, the lower and upper bounds take values 1.678 and 2, respectively.

5 Experimental Results

In this section we explore some empirical properties of the introduced indices using synthetic data. We generated four groups of 100 random $10 \times 10 \times 10$ contexts with densities 0.1, 0.2, 0.4, 0.6. The features of the data are given in Figure 2.

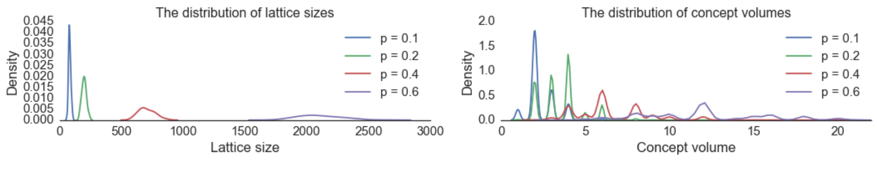

Fig. 2. Parameters of lattices constructed on $10 \times 10 \times 10$ contexts.

The choice of a subset of indices for data exploration can be motivated by the following properties: the indices should be pairwise uncorrelated (to avoid biased results when combining indices) and efficiently computable (if possible). The density function of an index may be a multimodal mixture of two or more distributions. In this case one needs a special justification for the choice of a threshold value separating two distributions.

We consider Pearson's correlation between all pairs of stability indices and cardinalities of sets that comprise a triadic concept (extent, intent, modus). In Figure 3 the values of the coefficient are represented. The sizes of the extent, intent and modus are denoted by $|A_1|$, $|A_2|$, $|A_3|$, respectively. The sizes of dyadic subcontexts are denoted in a similar way. The corresponding stability estimates are referred to by the *log* prefix.

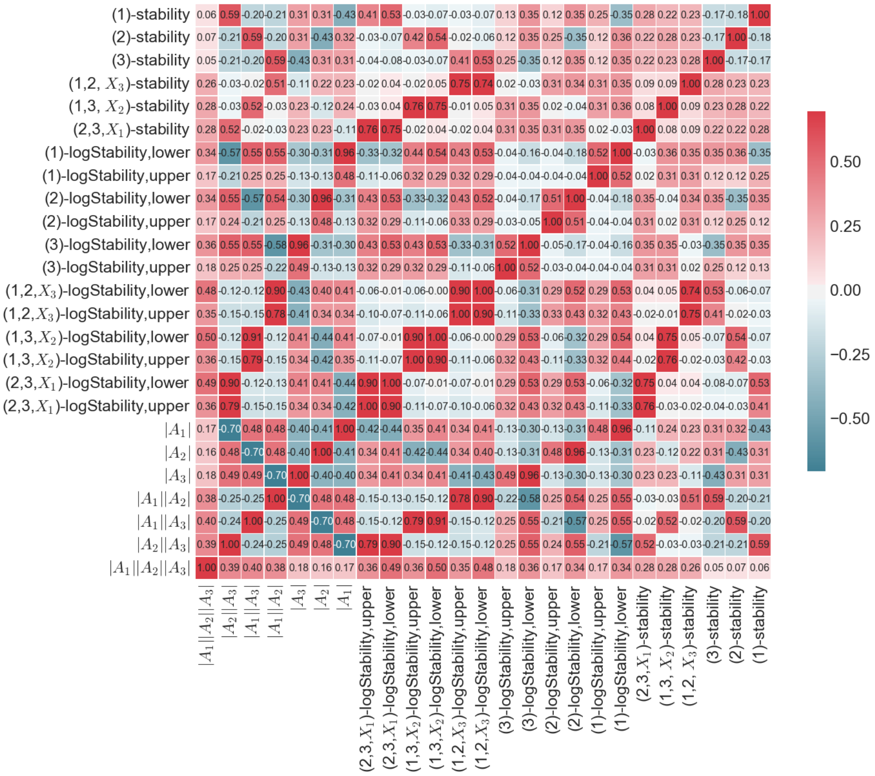

Fig. 3. The Pearson's correlation coefficient among 100 contexts with the density 0.4

As can be seen from Figure 3, there is a correlation between (i)-stability and $|A_j| \cdot |A_k|$. The index of (i)-stability correlates less strongly with the estimates of (j, k, X_i)-stability and the size of the set A_i. These types of correlation become stronger as the density of a context increases. In fact, these indices can be replaced by the size of a particular set in the case of a dense context. A correlation is observed between (i, j, X_k)-stability and its estimates, a less strong correlation is observed between (i, j, X_k)-stability and $|A_i| \cdot |A_j|$. It is important to note that the strong correlation between (i, j, X_k)-stability and its estimates, as well as very small correlation between (i, j, X_k)-stability and estimates of (k)-stability remains the same with different context densities. The pairwise correlation between stability indices is weak, hence it is preferable to use these indices together.

For selecting interesting concepts based on values of an index it is important to choose a correct threshold. This choice can be based on the distribution of index values. Figure 4 shows that the distribution of values (2)-stability and $(1, 3, X_k)$-stability (other (i)-stabilities and (i, j, X_k)-stabilities have similar distributions). The distribution of (i)-stability values allows us to identify a threshold easily, since some picks exist in the distribution, while for (i, j, X_k)-stability the distribution varies from density to density, in case of a dense context it motivates further study of the index and the way one selects thresholds for it. For values of the lower bound of stability estimates the modes of the distribution become less distinct or the distribution becomes unimodal (Figures 5,6). The upper bound for (i)-stability estimate (or (i, j, X_k)) in most cases corresponds to $|A_i|$ (or $|A_i| \cdot |A_j|$), since the closure of a superset of $A_j \times A_k$ (or A_k) results in the empty set.

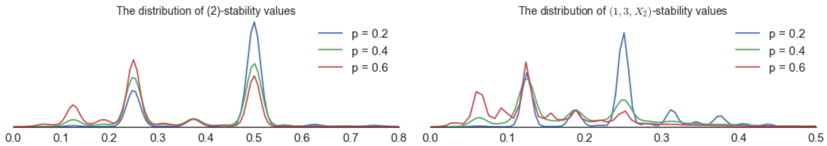

Fig. 4. The distribution of values of (2)-stability and $(1, 3, X_2)$-stability

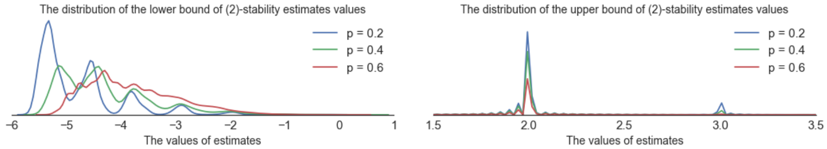

Fig. 5. The values distribution of (2)-stability estimates

The lower bound of (i)-stability is also strongly correlated with the size of set i. This is due to the fact that a big size of the set i leads to larger difference between the sizes of A_i and $(X_j \times X_k)^{(i)}$, where $X_j \times X_k$ is a superset of $A_j \times A_k$, and the sum under logarithm. The estimate of (i, j, X_k)-stability are correlated

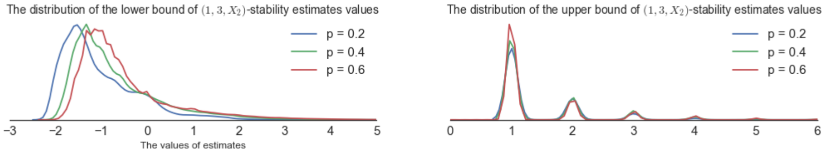

Fig. 6. The values distribution of (i, j, X_k)-stability estimates

with the corresponding indices. There is roughly the same correlation between the estimate and the value $|A_i| \cdot |A_j|$, which results from the bigger difference between $|A_i| \cdot |A_j|$ and $|X_i| \cdot |X_j|$, which correspond to a superset of A_k. The upper and lower estimates of (i, j, X_k)-stability also correlate, in this case, the correlation could be related to set-difference between the set $A_i \times A_j$ and the volume of the rectangular subarea of $X_i \times X_k$ for the corresponding superset of A_k.

It is noteworthy that the calculation of stability estimates in practice could take more time then the stability calculation itself. It is typical for (i)-stability, where the number $2^{|A_i|}$ is lower then the number of all possible subsets obtained by adding elements from $K_j \times K_k \setminus A_j \times A_k$.

6 Conclusion

In this paper we have introduced two stability indices for triadic concepts, based on two derivation operators, and studied their empirical behavior. We have proposed to compute stability using two derivation operators. We have studied correlation of stability indices and their distributions, which is important in practical data analysis. As it was shown, the introduced stability indices are not pairwise correlated and therefore can be used in some combinations for selecting interesting concepts. Moreover, (i)-stability correlates with $|A_i|$(for dense contexts) and $|A_j||A_k|$, and hence these indices should not be combined together.

The values of (i)-stability for all concepts are characterized by the presence of groups of values with high frequency, which facilitates selection of interesting concepts based on threshold values, while the distribution of (i, j, X_k)-stability does not give clearly defined groups of interesting concepts.

We have also introduced the estimates of stability indices, which correlate both with the corresponding stability indices and some of stability estimates. This is due to the fact that the estimates of (i)-stability (or (i, j, X_k)-stability) are based on the elements from $K_j \times K_k \setminus A_j \times A_k$ (or $K_k \setminus A_k$). Hence, the choice between stability and its estimates must be guided by the comparison of the sizes of sets involved in calculation, e.g. in the case of (i)-stability the number of subsets $2^{|A_i|}$, most probably, will be less then the number of elements in $K_j \times K_k \setminus A_j \times A_k$.

The proposed indices characterize triconcepts differently, in general they do not agree in the top-n selected concepts, which allow us to use either their combination to set up the strictest selection criteria, or to take some of them depending on the meaning behind a stability index.

References

1. Lehmann, F., Wille, R.: A triadic approach to formal concept analysis. In: Conceptual Structures: Applications, Implementation and Theory: Third International Conference on Conceptual Structures, ICCS '95 Santa Cruz, CA, USA, August 14–18, 1995 Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg (1995) 32–43
2. Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G.: Trias–an algorithm for mining iceberg tri-lattices. In: Proceedings of the Sixth International Conference on Data Mining. ICDM '06, Washington, DC, USA, IEEE Computer Society (2006) 907–911
3. Ji, L., Tan, K.L., Tung, A.K.H.: Mining frequent closed cubes in 3d datasets. In: Proceedings of the 32Nd International Conference on Very Large Data Bases. VLDB '06, VLDB Endowment (2006) 811–822
4. Buzmakov, A., Kuznetsov, S.O., Napoli, A.: Scalable estimates of concept stability. In Glodeanu, C., Kaytoue, M., Sacarea, C., eds.: Formal Concept Analysis. Lecture Notes in Computer Science, Springer International Publishing (ICFCA, 2014) 157–172
5. Ganter, B., Wille, R.: Contextual attribute logic. In Tepfenhart, W., Cyre, W., eds.: Conceptual Structures: Standards and Practices. Volume 1640 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (1999) 377–388
6. Kuznetsov, S.O.: Stability as an estimate of degree of substantiation of hypotheses derived on the basis of operational similarity. Nauchn. Tekh. Inf., Ser. 2 (12) (1990) 21–29
7. Kuznetsov, S.O.: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence **49**(1-4) (2007) 101–115
8. Kuznetsov, S.O., Obiedkov, S., Roth, C.: Reducing the representation complexity of lattice-based taxonomies. In: Conceptual Structures: Knowledge Architectures for Smart Applications. Springer Berlin Heidelberg (2007) 241–254
9. Babin, M.A., Kuznetsov, S.O.: Approximating concept stability. In Domenach, F., Ignatov, D., Poelmans, J., eds.: Formal Concept Analysis. Volume 7278 of Lecture Notes in Computer Science., Springer Berlin Heidelberg (2012) 7–15