

# Analysing Online Social Network Data with Biclustering and Triclustering

Dmitry Gnatyshak<sup>1</sup>, Dmitry I. Ignatov<sup>1</sup>, Alexander Semenov<sup>1</sup>, and Jonas Poelmans<sup>1,2</sup>

<sup>1</sup> National Research University Higher School of Economics, Russia  
dignatov@hse.ru

<http://www.hse.ru>

<sup>2</sup> Katholieke Universiteit Leuven, Belgium

**Abstract.** In this paper we propose two novel methods for analysing data collected from online social networks. In particular we will do analyses on Vkontake data (Russian online social network). Using biclustering we extract groups of users with similar interests and find communities of users which belong to similar groups. With triclustering we reveal users' interests as tags and use them to describe Vkontakte groups. After this social tagging process we can recommend to a particular user relevant groups to join or new friends from interesting groups which have a similar taste. We present some preliminary results and explain how we are going to apply these methods on massive data repositories.

**Keywords:** Formal Concept Analysis, Biclustering and Triclustering, Online Social Networks, Web 2.0 and social computing

## 1 Introduction

Recently the focus of social network analysis shifted from 1-mode networks, like friend-to-friend, to 2-mode [1,2,3], 3-mode [4,5,6] and even multimodal dynamic networks [7,8,9].

This interest is not only pure academic but caused by modern business requirements. Thus, every user of a social networking website has not only friends, but he also has specific profile features, e.g. he can belong to some groups of users, indicate his tastes or books he read etc. These profile attributes are able to describe the user's tastes, preferences, attitudes, which is highly relevant for business oriented social networking web sites owners. Finding bicomunities and tricommunities can help the networking site owners to analyze large groups of their users and adjust their services according to users' needs which may in the end result in financial or other benefits.

There is a large amount of network data that can be represented as bipartite or tripartite graphs. Standard techniques like "maximal bicliques search" return a huge number of patterns (in the worst case exponential w.r.t. the input size). Therefore we need some relaxation of the biclique notion and good interestingness measures for mining biclique communities.

Applied lattice theory provides us with a notion of formal concept [10] which is the same thing as a biclique; it is widely known in the social network analysis community (see, e.g. [11,12,13,14,15,16]).

A concept-based bicluster [17] is a scalable approximation of a formal concept (biclique). The advantages of concept-based biclustering are:

1. Less number of patterns to analyze;
2. Less computational time (polynomial vs exponential);
3. Manual tuning of bicluster (community) density threshold;
4. Tolerance to missing (object, attribute) pairs.

For analyzing three-mode network data like folksonomies [18] we also proposed a triclustering technique [19]. In this paper we describe a new pseudo-triclustering technique for tagging groups of users by their common interest. This approach differs from traditional triclustering methods because it relies on the extraction of biclusters from two separate object-attribute tables. Biclusters which are similar with respect to their extents are merged by taking the intersection of the extents. The intent of the first bicluster and the intent of the second bicluster become the intent and modus respectively of the newly obtained tricluster. Our approach was empirically validated on online social network data obtained from V Kontakte (<http://vk.com>).

The remainder of the paper is organized as follows. In section 2 we describe some key notions from Formal Concept Analysis. In section 3 we introduce a model for our new pseudo-triclustering approach. In section 4 we describe a dataset which is a sample of users, their groups and interests from the V Kontakte (<http://vk.com>) social networking web site. We present the results obtained during experiments on this dataset in Section 5. Section 6 concludes our paper and describes some interesting directions for future research.

## 2 Basic definitions

The *formal context* in FCA [10] is a triple  $\mathbb{K} = (G, M, I)$ , where  $G$  is a *set of objects*,  $M$  is a *set of attributes*, and the relation  $I \subseteq G \times M$  shows which object which attribute possesses. For any  $A \subseteq G$  and  $B \subseteq M$  one can define *Galois operators*:

$$\begin{aligned} A' &= \{m \in M \mid gIm \text{ for all } g \in A\}, \\ B' &= \{g \in G \mid gIm \text{ for all } m \in B\}. \end{aligned} \tag{1}$$

The operator  $''$  (applying the operator  $'$  twice) is a *closure operator*: it is idempotent ( $A'''' = A''$ ), monotonous ( $A \subseteq B$  implies  $A'' \subseteq B''$ ) and extensive ( $A \subseteq A''$ ). The set of objects  $A \subseteq G$  such that  $A'' = A$  is called *closed*. The same is for closed attributes sets, subsets of a set  $M$ . A couple  $(A, B)$  such that  $A \subseteq G$ ,  $B \subseteq M$ ,  $A' = B$  and  $B' = A$ , is called *formal concept* of a context  $K$ . The sets  $A$  and  $B$  are closed and called *extent* and *intent* of a formal concept  $(A, B)$

correspondingly. For the set of objects  $A$  the set of their common attributes  $A'$  describes the similarity of objects of the set  $A$ , and the closed set  $A''$  is a cluster of similar objects (with the set of common attributes  $A'$ ). The relation “to be a more general concept” is defined as follows:  $(A, B) \geq (C, D)$  iff  $A \subseteq C$ . The concepts of a formal context  $\mathbb{K} = (G, M, I)$  ordered by extensions inclusion form a lattice, which is called *concept lattice*. For its visualization the *line diagrams* (Hasse diagrams) can be used, i.e. cover graph of the relation “to be a more general concept”. In the worst case (Boolean lattice) the number of concepts is equal to  $2^{\{\min |G|, |M|\}}$ , thus, for large contexts, FCA can be used only if the data is sparse. Moreover, one can use different ways of reducing the number of formal concepts (choosing concepts by their stability index or extent size). The alternative approach is a relaxation of the definition of formal concept as a maximal rectangle in an object-attribute matrix which elements belong to the incidence relation. One of such relaxations is the notion of an object-attribute bicluster [17]. If  $(g, m) \in I$ , then  $(m', g')$  is called object-attribute bicluster with the density  $\rho(m', g') = |I \cap (m' \times g')| / (|m'| \cdot |g'|)$ .

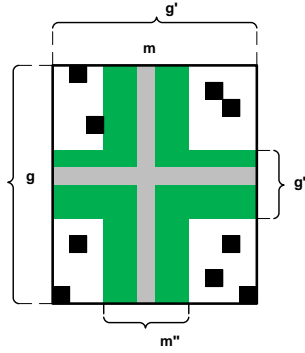


Fig. 1. OA-bicluster.

The main features of OA-biclusters are listed below:

1. For any bicluster  $(A, B) \subseteq 2^G \times 2^M$  it is true that  $0 \leq \rho(A, B) \leq 1$ .
2. OA-bicluster  $(m', g')$  is a formal concept iff  $\rho = 1$ .
3. If  $(m', g')$  is a bicluster, then  $(g'', g') \leq (m', m'')$ .

Let  $(A, B) \subseteq 2^G \times 2^M$  be a bicluster and  $\rho_{\min}$  be a non-negative real number such that  $0 \leq \rho_{\min} \leq 1$ , then  $(A, B)$  is called *dense*, if it fits the constraint  $\rho(A, B) \geq \rho_{\min}$ . The above mentioned properties show that OA-biclusters differ from formal concepts by the fact that they do not necessarily have unit density. Graphically it means that not all the cells of a bicluster must be filled by a cross (see fig. 1).

### 3 Model and algorithm description

Let  $\mathbb{K}_{UI} = (U, I, X \subseteq U \times I)$  be a formal context which describes what interest  $i \in I$  a particular user  $u \in U$  has. Similarly, let  $\mathbb{K}_{UG} = (U, G, Y \subseteq U \times G)$  be a formal context which indicates what group  $g \in G$  user  $u \in U$  belongs to.

We can find dense biclusters as (*users, interests*) pairs in  $\mathbb{K}_{UI}$  using the OA-biclustering algorithm which is described in [17]. These biclusters are groups of users who have similar interests. In the same way we can find communities of users, who belong to similar groups on the Vkontakte social network, as dense biclusters (*users, groups*).

By means of triclustering we can also reveal users' interests as tags which describe similar Vkontakte groups. So, by doing this we can solve the task of social tagging and recommend to a particular user relevant groups to join or interests to indicate on the page or new friends from interesting groups with similar tastes to follow.

To this end we need to mine a (formal) tricontext  $\mathbb{K}_{UIG} = (U, I, G, Z \subseteq U \times I \times G)$ , where  $(u, i, g)$  is in  $Z$  iff  $(u, i) \in X$  and  $(u, g) \in Y$ . A particular tricluster has a form  $T_k = (i^X \cap g^Y, u^X, u^Y)$  for every  $(u, g, i) \in Z$  with  $\frac{|i^X \cap g^Y|}{|i^X \cup g^Y|} \geq \Theta$ , where  $\Theta$  is a predefined threshold between 0 and 1. We can calculate the density of  $T_k$  directly, but it takes  $O(|U||I||G|)$  time in the worst case, so we prefer to define the quality of such tricluster by density of biclusters  $(g^Y, u^Y)$  and  $(i^X, u^X)$ . We propose to calculate this estimator as  $\hat{\rho}(T_k) = \frac{\rho(g^Y, u^Y) + \rho(i^X, u^X)}{2}$ ; it's obvious that  $0 \leq \hat{\rho} \leq 1$ . We have to note that the third component of a (pseudo)tricluster or triadic formal concept usually is called *modus*.

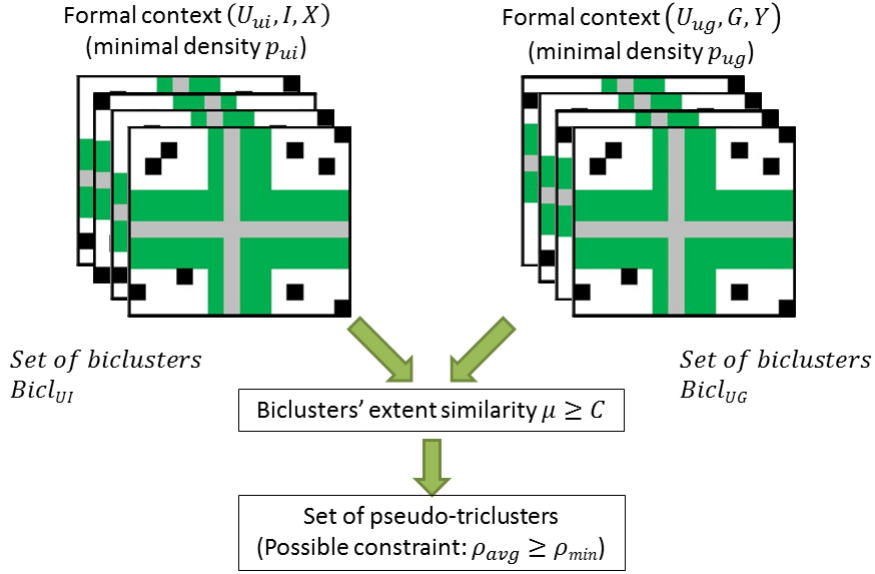
The algorithm scheme is displayed in Fig. 2

### 4 Data

For our experiments we collected a dataset from the Russian social networking site Vkontakte. Each entry consisted of the following fields: id, userid, gender, family status, birthdate, country, city, institute, interests, groups. This set was divided into 4 subsets based on the values of the institute field, namely students of two major technical universities and two universities focusing on humanities and sociology were considered: The Bauman Moscow State Technical University, Moscow Institute of Physics and Technology (MIPT), the Russian State University for Humanities (RSUH) and the Russian State Social University (RSSU). Then 2 formal contexts, users-interests and users-groups were created for each of these new subsets.

### 5 Experiments

We performed our experiments under the following setting: Intel Core i7-2600 system with 3.4 GHz and 8 GB RAM. For each of the created datasets the following experiment was conducted: first of all, two sets of biclusters using



**Fig. 2.** Pseudo-triclustering algorithm scheme

**Table 1.** Basic description of four data sets of large Russian universities.

	Bauman	MIPT	RSUH	RSSU
number of users	18542	4786	10266	12281
number of interests	8118	2593	5892	3733
number of groups	153985	46312	95619	102046

various minimal density constraints were generated, one for each formal context. Then the sets fulfilling the minimal density constraint of 0.5 were chosen, each pair of their biclusters was enumerated and the pairs with sufficient extents intersection ( $\mu$ ) were added to the corresponding pseudo-tricluster sets. This process was repeated for various values of  $\mu$ .

As it can be seen from the graphs and the tables, the majority of pseudo-triclusters had  $\mu$  value of 0.3 (or, to be more precise, 0.33). In this series of experiments we didn't reveal manually any interests which are particular for certain universities, but the number of biclusters and pseudo-triclusters was relatively higher for Bauman State University. This is a direct consequence of the higher users' number and the diversity of their groups.

Some examples of obtained biclusters and triclusters with high values of density and similarity are presented below.

**Example 1.** Biclusters in the form  $(Users, Intersts)$ .

**Table 2.** Bicluster density distribution and elapsed time for different  $\rho_{min}$  thresholds (Bauman and MIPT universities).

$\rho$	Bauman				MIPT			
	UI		UG		UI		UG	
	Time, s	Number	Time, s	Number	Time, s	Number	Time, s	Number
0.0	9.188	8863	1874.458	248077	0.863	2492	109.012	46873
0.1	8.882	8331	1296.056	173786	0.827	2401	91.187	38226
0.2	8.497	6960	966.000	120075	0.780	2015	74.498	28391
0.3	8.006	5513	788.008	85227	0.761	1600	63.888	21152
0.4	7.700	4308	676.733	59179	0.705	1270	56.365	15306
0.5	7.536	3777	654.047	53877	0.668	1091	54.868	13828
0.6	7.324	2718	522.110	18586	0.670	775	44.850	5279
0.7	7.250	2409	511.711	15577	0.743	676	43.854	4399
0.8	7.217	2326	508.368	14855	0.663	654	43.526	4215
0.9	7.246	2314	507.983	14691	0.669	647	43.216	4157
1.0	7.236	2309	511.466	14654	0.669	647	43.434	4148

**Table 3.** Bicluster density distribution and elapsed time for different  $\rho_{min}$  thresholds (RSUH and RSSU universities).

$\rho$	RSUH				RSSU			
	UI		UG		UI		UG	
	Time, s	number	Time, s	number	Time, s	number	Time, s	number
0.0	3,958	5293	519.772	116882	2.588	4014	693.658	145086
0.1	3.763	4925	419.145	93219	2.450	3785	527.135	110964
0.2	3.656	4003	330.371	68709	2.369	3220	402.159	79802
0.3	3.361	3123	275.394	50650	2.284	2612	332.523	58321
0.4	3.252	2399	232.154	35434	2.184	2037	281.164	40657
0.5	3.189	2087	224.808	32578	2.179	1782	270.605	37244
0.6	3.075	1367	174.657	10877	2.159	1264	211.897	12908
0.7	3.007	1224	171.554	9171	2.084	1109	208.632	10957
0.8	3.032	1188	170.984	8742	2.121	1081	209.084	10503
0.9	2.985	1180	174.781	8649	2.096	1072	206.902	10422
1.0	3.057	1177	173.240	8635	2.086	1068	207.198	10408

**Table 4.** Number of similar biclusters and elapsed time for different  $\mu$  thresholds (four universities).

$\mu$	Bauman		MIPT		RSUH		RSSU	
	Time, s	Count	Time, s	Count	Time, s	Count	Time, s	Count
0.0	3353.426	230161	77.562	24852	256.801	35275	183.595	55338
0.1	76.758	10928	35.137	5969	62.736	5679	18.725	5582
0.2	80.647	8539	31.231	4908	58.695	5089	16.466	3641
0.3	77.956	6107	27.859	3770	53.789	3865	17.448	2772
0.4	60.929	31	2.060	12	9.890	14	13.585	12
0.5	66.709	24	2.327	10	9.353	14	12.776	10
0.6	57.803	22	2.147	8	11.352	14	12.268	10
0.7	68.361	18	2.333	8	10.778	12	13.819	4
0.8	70.948	18	2.256	8	9.489	12	13.725	4
0.9	65.527	18	1.942	8	10.769	12	11.705	4
1.0	65.991	18	1.971	8	10.763	12	13.263	4

- $\rho = 83, 33\%$ , generator pair:  $\{3609, home\}$ ,  
bicluster:  $(\{3609, 4566\}, \{family, work, home\})$
- $\rho = 83, 33\%$ , generator pair:  $\{30568, orthodox\ church\}$ ,  
bicluster:  $(\{25092, 30568\}, \{music, monastery, orthodox\ church\})$
- $\rho = 100\%$ , generator pair:  $\{4220, beauty\}$ ,  
bicluster:  $(\{1269, 4220, 5337, 20787\}, \{love, beauty\})$

E.g., the second bicluster can be read as users 25092 and 30568 have almost all “music”, “monastery”, “orthodox church” as common interests. The pair generator shows which pair (*user, interest*) was used to build a particular bicluster.

**Example 2.** Pseudo-triclusters in the form (*Users, Interests, Groups*).

Bicluster similarity  $\mu = 100\%$ , average density  $\hat{\rho} = 54, 92\%$ .

Users:  $\{16313, 24835\}$ ,

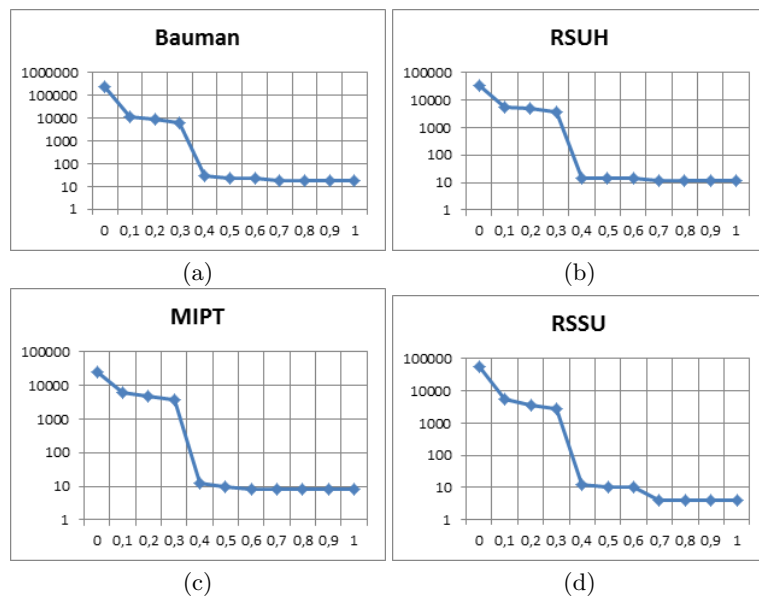
Interests:  $\{sleeping, painting, walking, tattoo, hamster, impressions\}$ ,

Groups:  $\{365, 457, 624, \dots, 17357688, 17365092\}$

This tricluster can be interpreted as a set of two users who have on average 55% of common interests and groups. The two corresponding biclusters have the same extents, i.e. people with almost all interests from the intent of this tricluster and people with almost all groups from the tricluster modus coincide.

## 6 Conclusions

The approach needs some improvements and fine tuning in order to increase the scalability and quality of the community finding process. We consider several directions for improvements: Strategies for approximate density calculation; Choosing good thresholds for  $n$ -clusters density and communities similarity; More sophisticated quality measures like recall and precision in Information



**Fig. 3.** Density bicluster distribution for the empirical data sets of four Russian universities. (a) Bauman State University (b) Russian State University for Humanities (c) Moscow Physical University (d) Russian State Social University

Retrieval; The proposed technique also needs comparison with other approaches like iceberg lattices ([20]), stable concepts ([21]), fault-tolerant concepts ([22]) and different  $n$ -clustering techniques from bioinformatics ([23], [24], etc.). We also claim that it is possible to obtain more dense pseudo-triclusters based on conventional formal concepts (even though it is expensive from a computational point of view). To validate the relevance of the extracted tricommunities expert feedback (e.g., validation by sociologist) is needed.

Finally, we conclude that it is possible to use our pseudo-triclustering method for tagging groups by interests in social networking sites and finding tricommunities. E.g., if we have found a dense pseudo-tricluster (*Users, Groups, Interests*) we can mark *Groups* by user interests from *Interests*. It also makes sense to use biclusters and triclusters for making recommendations. Missing pairs and triples seem to be good candidates to recommend the target user other potentially interesting users, groups and interests.

**Acknowledgments.** We would like to thank our colleagues Vincent Duquenne, Sergei Kuznetsov, Sergei Obiedkov, Camille Roth and Leonid Zhukov for their inspirational discussions, which directly or implicitly influenced this study.



## References

1. Latapy, M., Magnien, C., Vecchio, N.D.: Basic notions for the analysis of large two-mode networks. *Social Networks* **30**(1) (2008) 31 – 48
2. Liu, X., Murata, T.: Evaluating community structure in bipartite networks. In Elmagarmid, A.K., Agrawal, D., eds.: *SocialCom/PASSAT*, IEEE Computer Society (2010) 576–581
3. Opsahl, T.: Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks* **34** (2011) – (in press).
4. 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
5. Murata, T.: Detecting communities from tripartite networks. In Rappa, M., Jones, P., Freire, J., Chakrabarti, S., eds.: *WWW*, ACM (2010) 1159–1160
6. Ignatov, D.I., Kuznetsov, S.O., Magizov, R.A., Zhukov, L.E.: From Triconcepts to Triclusters. In: *Proceedings of the 13th international conference on Rough sets, fuzzy sets, data mining and granular computing. RSFDGrC'11*, Berlin, Heidelberg, Springer-Verlag (2011) 257–264
7. Roth, C.: Generalized preferential attachment: Towards realistic socio-semantic network models. In: *ISWC 4th Intl Semantic Web Conference, Workshop on Semantic Network Analysis*, Galway, Ireland,. Volume 171 of CEUR-WS Series (ISSN 1613-0073). (2005) 29–42
8. Roth, C., Cointet, J.P.: Social and semantic coevolution in knowledge networks. *Social Networks* **32** (2010) 16–29
9. Yavorsky, R.: Research Challenges of Dynamic Socio-Semantic Networks. In Ignatov, D., Poelmans, J., Kuznetsov, S., eds.: *CEUR Workshop proceedings Vol-757, CDUD'11 - Concept Discovery in Unstructured Data*. (2011) 119–122
10. Ganter, B., Wille, R.: *Formal Concept Analysis: Mathematical Foundations*. 1st edn. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1999)
11. Freeman, L.C., White, D.R.: Using galois lattices to represent network data. *Sociological Methodology* **23** (1993) 127–146
12. Freeman, L.C.: Cliques, galois lattices, and the structure of human social groups. *Social Networks* **18** (1996) 173–187
13. Duquenne, V.: Lattice analysis and the representation of handicap associations. *Social Networks* **18**(3) (1996) 217 – 230
14. White, D.R.: Statistical entailments and the galois lattice. *Social Networks* **18**(3) (1996) 201 – 215
15. Mohr, J.W., Duquenne, V.: The Duality of Culture and Practice: Poverty Relief in New York City, 1888-1917. *Theory and Society*, Special Double Issue on New Directions in Formalization and Historical Analysis **26**(2/3) (April-June 1997) 305–356
16. Roth, C., Obiedkov, S.A., Kourie, D.G.: Towards concise representation for taxonomies of epistemic communities. In Yahia, S.B., Nguifo, E.M., Belohlávek, R., eds.: *CLA*. Volume 4923 of *Lecture Notes in Computer Science*., Springer (2006) 240–255
17. Ignatov, D.I., Kaminskaya, A.Y., Kuznetsov, S., Magizov, R.A.: Method of Biclusterization Based on Object and Attribute Closures. In: *Proc. of 8-th international Conference on Intellectualization of Information Processing (IIP 2011)*. Cyprus, Paphos, October 17–24, MAKS Press (2010) 140 – 143 (in Russian).

18. Vander Wal, T.: Folksonomy Coinage and Definition (2007) <http://vanderwal.net/folksonomy.html> (accessed on 12.03.2012).
19. Ignatov, D.I., Kuznetsov, S.O., Magizov, R.A., Zhukov, L.E.: From triconcepts to triclusters. In: Proceedings of the 13th international conference on Rough sets, fuzzy sets, data mining and granular computing. RSFDGrC'11, Berlin, Heidelberg, Springer-Verlag (2011) 257–264
20. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with titanic. *Data & Knowledge Engineering* **42**(2) (August 2002) 189–222
21. Kuznetsov, S.O.: On stability of a formal concept. *Ann. Math. Artif. Intell.* **49**(1-4) (2007) 101–115
22. Besson, J., Robardet, C., Boulicaut, J.F.: Mining a new fault-tolerant pattern type as an alternative to formal concept discovery. In Scharfe, H., Hitzler, P., Ohrstrom, P., eds.: *Conceptual Structures: Inspiration and Application*. Volume 4068 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg (2006) 144–157
23. Zhao, L., Zaki, M.J.: Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data. SIGMOD '05, New York, NY, USA, ACM (2005) 694–705
24. Mirkin, B.G., Kramarenko, A.V.: Approximate bicluster and tricluster boxes in the analysis of binary data. In: Proceedings of the 13th international conference on Rough sets, fuzzy sets, data mining and granular computing. RSFDGrC'11, Berlin, Heidelberg, Springer-Verlag (2011) 248–256