Gaining Insight in Social Networks with Biclustering and Triclustering

Dmitry Gnatyshak[1], Dmitry I. Ignatov[1], Alexander Semenov[1], and Jonas Poelmans[1,2]

[1] National Research University Higher School of Economics, Russia
dignatov@hse.ru
http://www.hse.ru
[2] Katholieke Universiteit Leuven, Belgium

Abstract. We combine bi- and triclustering to analyse data collected from the Russian online social network Vkontakte. 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

Online social networks generate massive amounts of data which can become a valuable source for guiding Internet advertisement efforts. Each registered user has a network of friends as well as specific profile features. These profile features describe the user's tastes, preferences, the groups he or she belongs to, etc. Social network analysis is a popular research field in which methods are developed for analysing 1-mode networks, like friend-to-friend, 2-mode [1,2,3], 3-mode [4,5,6] and even multimodal dynamic networks [7,8,9]. We will focus on the subfield of bicommunity and tricommunity identification.

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:

N. Aseeva, E. Babkin, and O. Kozyrev (Eds.): BIR 2012, LNBIP 128, pp. 162–171, 2012.

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 [6]. 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 Vkontakte (`http://vk.com`).

The remainder of the paper is organized as follows. In section 2 we describe some key notions from Formal Concept Analysis and Biclustering. In section 3 we introduce the model for our new pseudo-triclustering approach. In section 4 we describe a dataset which consists of a sample of users, their groups and interests extracted from the Vkontakte (`http://vk.com`) social networking website. 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

2.1 Formal Concept Analysis

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 possesses which attribute. For any $A \subseteq G$ and $B \subseteq M$ one can define *Galois operators*:

$$A' = \{m \in M \mid gIm \text{ for all } g \in A\}, \qquad (1)$$
$$B' = \{g \in G \mid gIm \text{ for all } m \in B\}.$$

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 attribute sets, subsets of a set M. A couple (A, B) such that $A \subset G$, $B \subset M$, $A' = B$ and $B' = A$, is called a *formal concept* of a context $\mathbb{K}$. 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).

2.2 Biclustering

An 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 $\rho(m', g') = |I \cap (m' \times g')|/(|m'| \cdot |g'|)$.

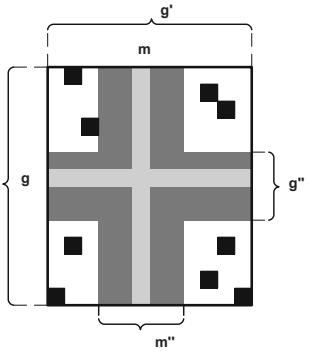

Fig. 1. OA-bicluster

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 ρ_{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). The rectangle in figure 1 depicts a bicluster extracted from an object-attribute table. The horizontal gray line corresponds to object g and contains only nonempty cells. The vertical gray line corresponds to attribute m and also contains only nonempty cells. By applying the Galois operator, as explained in section 2.1, one time to g we obtain all its attributes g'. By applying the Galois

operator twice to g we obtain all objects which have the same attributes as g. This is depicted in fig. 1 as g''. By applying the Galois operator twice to m we obtain all attributes which belong to the same objects as m. This is depicted in fig. 1 as m''. The white spaces indicate empty cells. The black dots indicate non-empty cells. Whereas a traditional formal concept would cover only the green and gray area, the bicluster also covers the white and black cells. This gives to OA-biclusters some desirable fault-tolerance properties.

Algorithm 1. Bicluster computation

Data: $K = (G, M, I)$ is a formal context, ρ_{min} is a threshold density value of bicluster density
Result: $B = \{(A_k, B_k) | (A_k, B_k)$ – bicluster$\}$

```
1  begin
2      Obj.Size = |G|
3      Attr.Size = |M|
4      B ⟵ ∅
5      for g ∈ G do
6          Obj[g] = g'
7      for m ∈ M do
8          Attr[m] = m'
9      for g ← 0 to |G| do
10         for m ∈ Obj[g] do
11             if ρ(Attr[m], Obj[g]) ≥ ρ_min then
12                 B.Add((Attr[m], Obj[g]))
```

For calculating biclusters fulfilling a minimal density requirement we need to perform several steps (see 1. The first step consists of applying the Galois operator to all objects in G and then to all attributes in M. Then all biclusters are enumerated in a sequential manner and only those fulfilling the minimal density requirement are retained.

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 $(users, interesets)$ 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 Θ 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 $\widehat{\rho}(T_k) = \frac{\rho(g^Y, u^Y) + \rho(i^X, u^X)}{2}$; it's obvious that $0 \leq \widehat{\rho} \leq 1$. 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. Intuitively, we start from two formal contexts from which we extract biclusters fulfilling the minimal density requirement set for each context separately. We then determine the similarity of biclusters' extents and in case of high similarity their intents are used to form a pseudo-tricluster. We select only those pseudo-triclusters whose average density is above a predefined threshold.

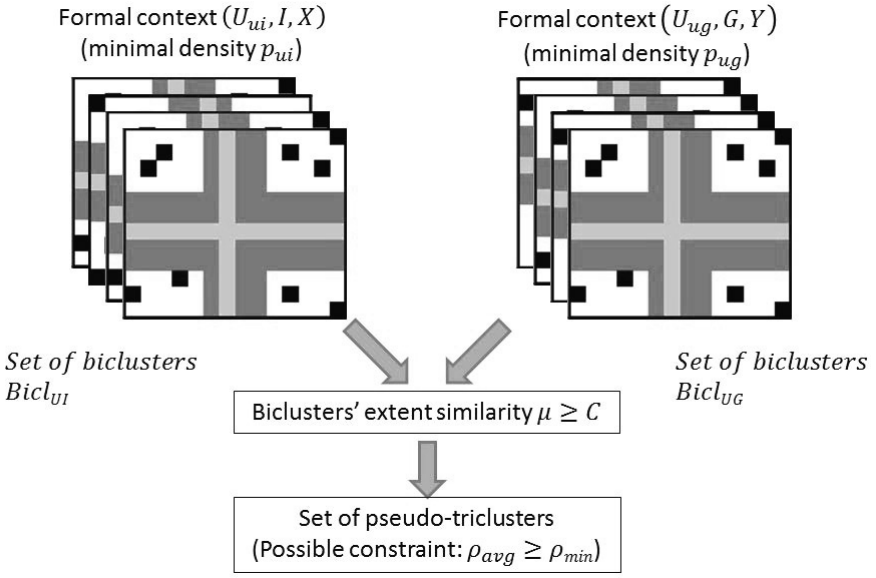

Fig. 2. Pseudo-triclustering algorithm scheme

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,

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

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 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 (μ) were added to the corresponding pseudo-tricluster sets. This process was repeated for various values of μ.

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

ρ	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 ρ_{min} thresholds (RSUH and RSSU universities)

ρ	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 μ thresholds (four universities)

μ	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

As it can be seen from the graphs and the tables, the majority of pseudo-triclusters had μ 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 $(Users, Intersts)$.

– $\rho = 83,33\%$, generator pair: $\{3609, home\}$,
bicluster: $(\{3609, 4566\}, \{family, work, home\})$

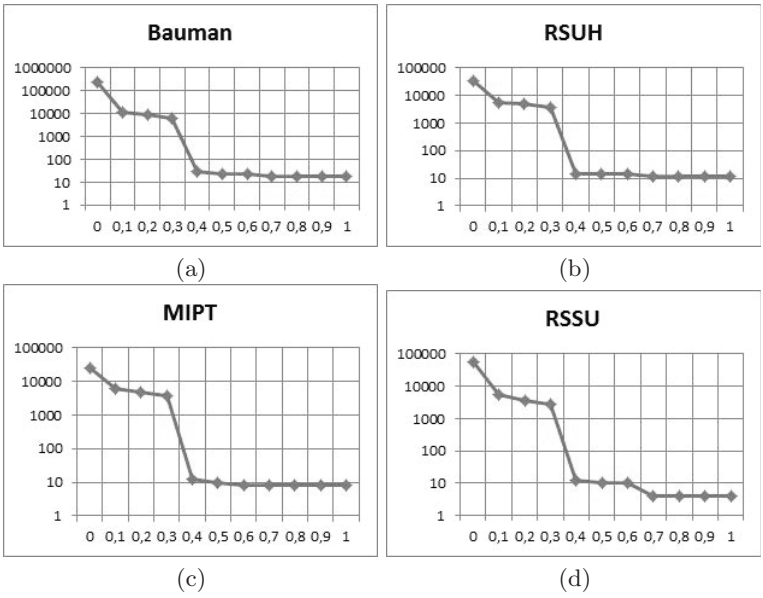

Fig. 3. Density bicluster distribution for the empirical data sets of four Russian universities. (a) Bauman State University (b) Russian State University for Humanities (c) Moscow Physical University (d) Russian State Social University

- $\rho = 83,33\%$, 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 bicuster 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 $(Users, Intersts, Groups)$.

Bicluster similarity $\mu = 100\%$, average density $\widehat{\rho} = 54,92\%$.
Users: $\{16313, 24835\}$,
Interests: $\{sleeping, painting, walking, tattoo, hamster, impressions\}$,
Groups: $\{365, 457, 624, \ldots, 17357688, 17365092\}$

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 Retrieval ([19]); 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 exctracted 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-trciluster ($Users$, $Groups$, $Interests$) 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. The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics in 2012 and in the Laboratory of Intelligent Systems and Structural Analysis.

References

1. Latapy, M., Magnien, C., Vecchio, N.D.: Basic notions for the analysis of large two-mode networks. Social Networks 30(1), 31–48 (2008)
2. Liu, X., Murata, T.: Evaluating community structure in bipartite networks. In: Elmagarmid, A.K., Agrawal, D. (eds.) SocialCom/PASSAT, pp. 576–581. IEEE Computer Society (2010)
3. Opsahl, T.: Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. Social Networks 34 (2011) ISSN: 0378-8733, http://www.sciencedirect.com/science/article/pii/S0378873311000360, doi:10.1016/j.socnet.2011.07.001
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 2006, pp. 907–911. IEEE Computer Society, Washington, DC (2006)
5. Murata, T.: Detecting communities from tripartite networks. In: Rappa, M., Jones, P., Freire, J., Chakrabarti, S. (eds.) WWW, pp. 1159–1160. ACM (2010)
6. Ignatov, D.I., Kuznetsov, S.O., Magizov, R.A., Zhukov, L.E.: From Triconcepts to Triclusters. In: Kuznetsov, S.O., Ślęzak, D., Hepting, D.H., Mirkin, B.G. (eds.) RSFDGrC 2011. LNCS, vol. 6743, pp. 257–264. Springer, Heidelberg (2011)
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. CEUR-WS Series, vol. 171, pp. 29–42 (2005) ISSN 1613-0073

8. Roth, C., Cointet, J.P.: Social and semantic coevolution in knowledge networks. Social Networks 32, 16–29 (2010)
9. Yavorsky, R.: Research Challenges of Dynamic Socio-Semantic Networks. In: Ignatov, D., Poelmans, J., Kuznetsov, S. (eds.) CDUD 2011 - Concept Discovery in Unstructured Data. CEUR Workshop proceedings, vol. 757, pp. 119–122 (2011)
10. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations, 1st edn. Springer-Verlag New York, Inc., Secaucus (1999)
11. Freeman, L.C., White, D.R.: Using galois lattices to represent network data. Sociological Methodology 23, 127–146 (1993)
12. Freeman, L.C.: Cliques, galois lattices, and the structure of human social groups. Social Networks 18, 173–187 (1996)
13. Duquenne, V.: Lattice analysis and the representation of handicap associations. Social Networks 18(3), 217–230 (1996)
14. White, D.R.: Statistical entailments and the galois lattice. Social Networks 18(3), 201–215 (1996)
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), 305–356 (1997)
16. Roth, C., Obiedkov, S., Kourie, D.G.: Towards Concise Representation for Taxonomies of Epistemic Communities. In: Yahia, S.B., Nguifo, E.M., Belohlavek, R. (eds.) CLA 2006. LNCS (LNAI), vol. 4923, pp. 240–255. Springer, Heidelberg (2008)
17. Ignatov, D.I., Kaminskaya, A.Y., Kuznetsov, S., Magizov, R.A.: Method of Biclusterzation Based on Object and Attribute Closures. In: Proc. of 8th International Conference on Intellectualization of Information Processing (IIP 2011), Cyprus, Paphos, October 17-24, pp. 140–143. MAKS Press (2010) (in Russian)
18. Vander Wal, T.: Folksonomy Coinage and Definition (2007), `http://vanderwal.net/folksonomy.html` (accessed on March 12, 2012)
19. Poelmans, J., Ignatov, D.I., Viaene, S., Dedene, G., Kuznetsov, S.O.: Text Mining Scientific Papers: A Survey on FCA-Based Information Retrieval Research. In: Perner, P. (ed.) ICDM 2012. LNCS (LNAI), vol. 7377, pp. 273–287. Springer, Heidelberg (2012)
20. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with titanic. Data & Knowledge Engineering 42(2), 189–222 (2002)
21. Kuznetsov, S.O.: On stability of a formal concept. Ann. Math. Artif. Intell. 49(1-4), 101–115 (2007)
22. Besson, J., Robardet, C., Boulicaut, J.-F.: Mining a New Fault-Tolerant Pattern Type as an Alternative to Formal Concept Discovery. In: Schärfe, H., Hitzler, P., Øhrstrøm, P. (eds.) ICCS 2006. LNCS (LNAI), vol. 4068, pp. 144–157. Springer, Heidelberg (2006)
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 2005, pp. 694–705. ACM, New York (2005)
24. Mirkin, B.G., Kramarenko, A.V.: Approximate Bicluster and Tricluster Boxes in the Analysis of Binary Data. In: Kuznetsov, S.O., Ślęzak, D., Hepting, D.H., Mirkin, B.G. (eds.) RSFDGrC 2011. LNCS, vol. 6743, pp. 248–256. Springer, Heidelberg (2011)