# Visual Analytics in FCA-based Clustering

Yury Kashnitsky

Higher School of Economics, Moscow, Russia ykashnitsky@hse.ru

**Abstract.** Visual analytics is a subdomain of data analysis which combines both human and machine analytical abilities and is applied mostly in decision-making and data mining tasks. Triclustering, based on Formal Concept Analysis (FCA), was developed to detect groups of objects with similar properties under similar conditions. It is used in Social Network Analysis (SNA) and is a basis for certain types of recommender systems. The problem of triclustering algorithms is that they do not always produce meaningful clusters. This article describes a specific triclustering algorithm and a prototype of a visual analytics platform for working with obtained clusters. This tool is designed as a testing frameworkis and is intended to help an analyst to grasp the results of triclustering and recommender algorithms, and to make decisions on meaningfulness of certain triclusters and recommendations.

**Keywords:** visual analytics, formal concept analysis, triclustering, social network analysis.

## 1 Introduction

Classical Formal Concept Analysis (FCA) deals with data which describe a relationship between a set of objects and a set of attributes and provides methods to derive a concept hierarchy or formal ontology in them [1]. FCA is a powerful tool for revealing dependencies in data and is commonly applied to data mining (in particular, text mining), machine learning, knowledge management, semantic webs, software development, and biology.

As a natural extension of FCA, Triadic Concept Analysis (TCA) manages triadic data in a form of objects, their attributes, and conditions under which these objects have certain attributes [2]. A common example is a social network analysis with a context including users (objects), events they take part in (attributes) and interests (which might be regarded as conditions under which a user participates in a certain event).

As the task of finding all concepts or triconcepts is computationally challenging, certain relaxations of these terms have been introduced: biclusters [3] and triclusters [4]. Here we address triclusters, i.e. combinations of sets of objects, their attributes, and conditions where not every object must have each attribute. Triclustering provides an output in the form of object clusters with similar attributes under similar conditions. Therefore, it is applied to mining

users with common interests, applicants with similar competences or books labelled by close tags [5], [6]. Triclustering is also a basis for a certain type of recommender systems [7], [8].

Visual analytics is an increasingly popular branch of Computer Science which combines both human and computer qualities to solve a range of problems that might lay beyond the power of man or machine separately. Actually, it is a subdomain of data analysis focusing on decision-making through data preprocessing, data mining and interactive user interfaces. For instance, Siemens PLM software allows developers to collect, process, visualize report data in the 3D engineering environment, and make real-time decisions in the process of developing new vehicles. The same method is used in situational and decision-making centres, in nuclear power energetics, and in crime investigations.

In this paper, we explore these topics and describe a framework which uses visual analytics to solve some problems in FCA.

## 2 Visual analytics

### 2.1 Definition and specificity

Generalizing and selecting crucial aspects of various definitions of visual analytics [9], [10], here we propose the following one:

*Visual analytics* is a subdomain of data analysis focusing on analytical reasoning on the basis of interactive user interfaces in process of data mining, data preprocessing, knowledge representation, discovering dependencies, and decision-making.

Let us further consider core peculiarities of visual analytics and the tasks it is designed to solve: [11]

1. Visual analytics usually deals with complicated problems with big amounts of data requiring both human and machine resources.
2. The final goal of visual analysis is to enable users to obtain deep insight in problems to be solved which might include processing of large amounts of data from various sources. For this purpose visual analytics combines both human and technological resources. On one hand, data mining and statistics are the driving force of any automatic data analysis. On the other hand, human brain's aptitude for information perception and discovering dependencies in data complies to machine techniques and thus turns visual analytics into a promising sphere for further development.
3. In its development, visual analytics fosters in its turn the development of data mining, data representation and visualization, and analytical reporting.
4. Visual analytics also deals with human cognition, information perception, Computer Science, interactive and graphical design.
5. Visual analytics combines methods of information visualization and graphical data representation where visualization fosters human perception by the following means:
   (a) Enlarging data resources capacities makes user memorize less

(b) Reducing search, such as by representing a large amount of data in small space
(c) Enhancing recognition of patterns, such as when information is organized in space by its time relationships
(d) Supporting easy relationship inference
(e) Monitoring large amounts of potential events
(f) Providing techniques for dynamic data monitoring

### 2.2 Siemems

Siemens uses visual analytics techniques in its product lifecycle management (PLM) software enabling developers to collect, process, visualize report data in the 3D engineering environment, and make real-time decisions in the process of developing new vehicles. [1]

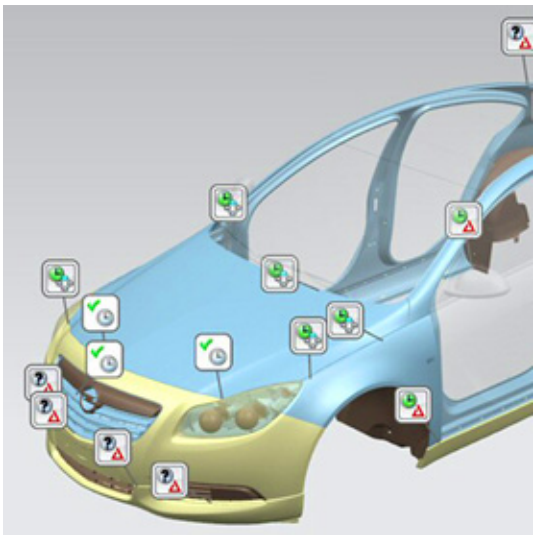

Fig. 1. One of development stages with Siemens PLM Software

The crucial point is that this system allows real-time visual interaction. This speeds up the processes of testing production for meeting given criteria, and eliminating product quality problems.

### 2.3 Supernova modelling

A highly powerful implementation of visual analytics paradigm was fulfilled by astrophysicists in Terascale Supernova Initiative (TSI) project. [2] The goal of

[1] http://www.plm.automation.siemens.com
[2] science.energy.gov/∼/media/ascr/ascac/pdf/meetings/mar03/Mezzacappa.pdf

the project is to give scientists from various fields access to powerful computation resources in order to produce knowledge in the sphere of fundamental science. In particular, the question of supernova birth was studied which encompassed 3D turbulence, gravitation and magnetic field modelling. The scale of the investigation was impressive - the modelling resulted in terabytes of data. The analysis of such amount of data lays beyond human power but combining human and machine capabilities allowed to make some inferences from all the bulk of data.

## 3 Formal Concept Analysis and OA-biclustering

### 3.1 Main definitions

A *formal context* in FCA is a triple $K = (G, M, I)$ where $G$ is a set of objects, $M$ is a set of attributes, and the binary relation $I \subseteq G \times M$ shows which object possesses which attribute. $gIm$ denotes that object $g$ has attribute $m$. For subsets of objects and attributes $A \subseteq G$ and $B \subseteq M$ *Galois operators* are defined as follows:

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

A pair $(A, B)$ such that $A \subset G, B \subset M, A' = B$ and $B' = A$, is called a *formal concept* of a context $K$. The sets $A$ and $B$ are closed and called the *extent* and the *intent* of a formal concept $(A, B)$ respectively. 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 number of formal concepts of a context $K = (G, M, I)$ can be quite large ($2^{min\{|G|,|M|\}}$ in the worst case), and the problem of computing this number is #P-complete [12]. There exist some ways to reduce the number of formal concepts, for instance, choosing concepts by stability, index or extent size [13].

An alternative way is to make a relaxation of the definition of a formal concept. One of them is an OA-bicluster [3].
If $(g, m) \in I$ , then $(m', g')$ is called an *object-attribute bicluster* with the *density*

$$\rho(m', g') = \frac{|I \bigcap (m' \times g')|}{(|m'||g'|)}.$$

Bicluster density represents the percent of object-attribute pairs from the initial context in a certain bicluster.

Here are the main properties of OA-biclusters:

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

A bicluster $(A, B)$ is called *dense* if its density is greater than a predefined minimum threshold, i.e. $\rho((A, B)) \geq \rho_{min}$. The above mentioned properties show that OA-biclusters differ from formal concepts since unit density is not required. Below follows an illustrative example for triconcepts and triclusters.

## 4 Triadic FCA and OAC-triclustering

As a solution for three-way data in FCA, Triadic Concept Analysis (TCA) was introduced [2].

A triadic context $K = (G, M, B, I)$ consists of sets $G$ (objects), $M$ (attributes), $B$ (conditions), and ternary relation $I \subseteq G \times M \times B$. An incidence $(g, m, b) \in I$ shows that the object $g$ has the attribute $m$ under condition $b$.

We denote a triadic context by $(X_1, X_2, X_3, I)$. A triadic context $K = (X_1, X_2, X_3, I)$ gives rise to the following dyadic contexts:

$$K^{(1)} = (X_1, X_2 \times X_3, I^{(1)}),$$
$$K^{(2)} = (X_2, X_3 \times X_1, I^{(2)}),$$
$$K^{(3)} = (X_3, X_1 \times X_2, I^{(3)}),$$

where $gI^{(1)}(m, b) \Leftrightarrow mI^{(1)}(g, b) \Leftrightarrow bI^{(1)}(g, m) \Leftrightarrow (g, m, b) \in I$.

The derivation operators (or prime operators) induced by $K^{(i)}$ are denoted by $(.)^{(i)}$. For each induced dyadic context we have two kinds of derivation operators. That is, for $\{i, j, k\} = \{1, 2, 3\}$ with $j < k$ and for $Z \subseteq X_i$ and $W \subseteq X_j \times X_k$ , the (i)-derivation operators are defined by:

$Z \to Z^{(i)} = \{(x_j, x_k) \in X_j \times X_k \mid x_i, x_j, x_k$ are related by $I$ for all $x_i \in Z\}$,

$W \to W^{(i)} = \{x_i \in X_i \mid x_i, x_j, x_k$ are related by $I$ for all $(x_j, x_k) \in W\}$

A *triadic concept* of a triadic context $K = (G, M, B, I)$ is a triple $(A_1, A_2, A_3)$ of $A_1 \subseteq X_1$, $A_2 \subseteq X_2$, $A_3 \subseteq X_3$ such that for every $\{i, j, k\} = \{1, 2, 3\}$ with $j < k$ we have $A_i^{(i)} = (A_j \times A_k)$.

$A_1, A_2$ and $A_3$ are called the *extent*, the *intent* and the *modus* of $(A_1, A_2, A_3)$.

A set $T = ((m, b)', (g, b)', (g, m)')$ for a triple $(g, m, b) \in I$ is called an *OAC-tricluster* (or object-attribute-condition tricluster or just tricluster) based on prime operators. Here

$$(g, m)' = \{b \mid (g, m, b) \in I\},$$
$$(g, b)' = \{m \mid (g, m, b) \in I\},$$
$$(m, b)' = \{g \mid (g, m, b) \in I\}.$$

The *density* of a tricluster $(A, B, C)$ of a triadic context $K = (G, M, B, I)$ is given by the fraction of all triples of $I$ in the tricluster, that is
$\rho(A, B, C) = \frac{|I \bigcap A \times B \times C|}{|A||B||C|}$.

The tricluster $T = (A, B, C)$ is called *dense* if its density is greater than a predefined minimum threshold, i.e. $\rho(T) \geq \rho_{min}$. Just similarly to biclusters, triclusters have the following properties:

1. For every triconcept $(A, B, C)$ of a triadic context $K = (G, M, B, I)$ with nonempty sets $A, B$ and $C$ we have $\rho(A, B, C) = 1$,
2. For every tricluster $(A, B, C)$ of a triadic context $K = (G, M, B, I)$ with nonempty sets $A, B$ and $C$ we have $0 \leq \rho(A, B, C) \leq 1$.

### 4.1 Example

Let us consider a sample context $K = (U, I, S, Y)$, where $U$ = {Ed, Leo, Max} is a set of users, $I$ = {soccer, hockey} — their interests, $S$ = {soccer.com, nhl.com, fifa.com, hockeycanada.ca} — sites they have added to bookmarks, $Y \subseteq U \times I \times S$ is a ternary relation between $U, I, S$ which can be expressed by Table 1:

| | $i_1$ | $i_2$ |
|---|---|---|
| $u_1$ | x | x |
| $u_2$ | x | x |
| $u_3$ | x | x |

| | $s_1$ | $s_2$ | $s_3$ | $s_4$ |
|---|---|---|---|---|
| $u_1$ | x | x | x | x |
| $u_2$ | x | x | x | |
| $u_3$ | x | x | x | x |

| | $s_1$ | $s_2$ | $s_3$ | $s_4$ |
|---|---|---|---|---|
| $i_1$ | x | | x | |
| $i_2$ | | x | | x |

Table 1. Sample context. Designations: $u_1$ - Ed, $u_2$ - Leo, $u_3$ - Max, $i_1$ - soccer, $i_2$ - hockey, $s_1$ - soccer.com, $s_2$ - nhl.com, $s_3$ - fifa.com, $s_4$ - hockeycanada.ca.

Here, generally, we have $|U||I||S| = 24$ triples to analyze. But actually, this number is reduced to 11, as there are lots of void triples present.

Actually, users Ed, Leo and Max share the same interests and almost the same sites (all the difference is that Leo has not bookmarked hockeycanada.ca). The idea of clustering here is presented by a tricluster $T = (\{u_1, u_2, u_3\}, \{i_1, i_2\}, \{s_1, s_2, s_3, s_4\})$ with density $\rho = 11/24 \cong 0.46$.
It is just one pattern to analyze instead of 11 in case of triples.

## 5 Implemented algorithms

The algorithms, described below, were implemented in Python 2.7.3 on a 2-processor machine (Core i3-370M, 2.4 HGz) with 3.87 GB RAM. One can find a description of testing procedure for these algorithms in [14] and [15].

### 5.1 OAC-prime triclustering algorithm

The hard core of the algorithm is quite simple: for all incidences $(g, m, b) \in I$ for a triadic context $K = (G, M, B, I)$ we build a tricluster $T = ((m, b)', (g, b)', (g, m)')$. If a tricluster is unique and its density exceeds a predefined minimum threshold then it is added to an array of triclusters. A pseudocode of algorithm for OAC-triclustering based on prime operators is presented below.

**Algorithm 1** OAC-triclustering based on prime operators

**Input:** $K = (G, M, B)$ - tricontext,
$\rho_{min}$ - density threshold
**Output:** $Tdic = \{X_1, X_2, X_3\}$ — a tricluster dictionary. $X_1 \subseteq G, X_2 \subseteq M, X_3 \subseteq B$

```
for (g, m, b) ∈ I do
    T = ((m, b)', (g, b)', (g, m)')
    HashKey = hash(T)
    if HashKey ∉ Tdic.keys() and ρ(T) ≥ ρ_min then
        Tdic[hashKey] = T
    end if
end for
```

### 5.2 Recommender algorithm based on triclustering

**Algorithm 2** Recommender algorithm

**Input:** $K = (U, T, R, I)$ - tricontext, $Tr$ - a set of triclusters
**Output:** $Tag_{rec}, Res_{rec}$ - sets of recommended tags and resources

```
for u ∈ U do
    for i = 1,...,|Tr| do
        sim_u(Tr_i) = 1/2 ( |R_u ∩ R_Tr_i| / |R_u ∪ R_Tr_i| + |T_u ∩ T_Tr_i| / |T_u ∪ T_Tr_i| )
        Tr_best = argmax(sim_u(Tr_i))
        Tag_rec[i] = T_Tr_best \ T_u
        Res_rec[i] = R_Tr_best \ R_u
    end for
end for
```

The recommender algorithm applied to sets of a tricontext is analogous to the one described in [7]. It takes as an input a context of three sets (objects, attributes, conditions), and the set of triclusters obtained as a result of the OAC-prime triclustering algorithm. For each user among all triclusters the one most similar to triples with this user is selected. The similarity of triclusters and triples is defined by function $sim_u(Tr_i)$. The algorithm returns sets $Tag_{rec}, Res_{rec}$ - tag and resource recommendations for all users.

## 6 The challenge and visual tricluster analysis framework

The challenge of the problem of triclustering (as of clustering on the whole) is to output meaningful, well-interpreted clusters. Actually, the term "meaningful" is not formally defined and is used by people to express their own subjective opinion on how well the task of clustering is solved, i.e. how similar the objects

in same clusters are, how distant - in different ones, how it corresponds to real world problems etc. Therefore, here an expert opinion might be useful, and a prototype of a visual analytics framework, described below, provides visual feedback for expert, and gives him ability to explore clusters in details.

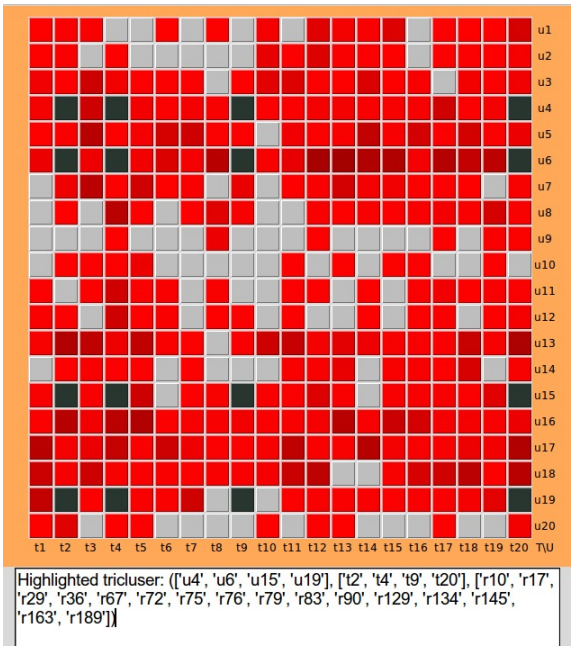


Fig. 2. Highlighting a largest tricluster for a user-tag pair $(u6, t4)$

In figure 2, we can see a map of triclusters produced by algorithm 1 for a context of 20 users, 20 tags, and 200 resources. The map is projected on the User-Tag plane. The more a certain user-tag pair is presented in triclusters the darker the corresponding square. A user-tag pair $(u6, t12)$, for instance, is included in 73 triclusters (a dark red square) while $(u5, t9)$ - just in 1 (a red square), and no triclusters have a pair $(u9, t10)$ (a grey one).

All triclusters including a certain user-tag pair can be listed by clicking on the "Triclusters" menu label. Similarly, triconcepts can be listed. One can also highlight the biggest tricluster with a certain user-tag pair or output all triclusters of the initial context ordered by density. Moreover, through the "Recommend attributes" context menu option an analyst can depict the results of recommender algorithm for a certain user (in this case, to show recommended tags).

The tool is intended to help an analyst to grasp the results of triclustering and recommender algorithms, and to make decisions on meaningfulness of certain triclusters and recommendations. The map helps the expert to quickly detect the concentrated regions (dark squares) and visualize dense triclusters including the

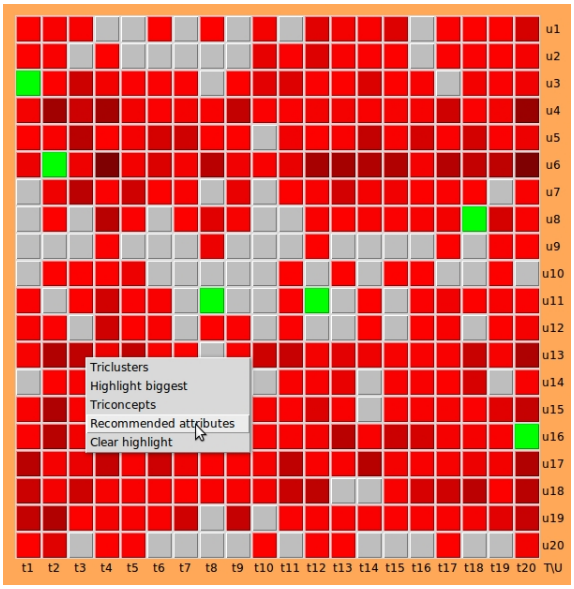


Fig. 3. Recommended tags for several users

corresponding triples. Further, it helps to make the decision whether the selected dense tricluster is meaningful or not, i.e. if it really combines similar users, tags, and resources.

## 7 Further work

There are several important issues to be regarded:

1. Limited human contribution: human contribution to triclustering in this visual analytics approach is limited and might only reach some hundreds of decisions on certain triclusters (less plausible, a thousand). Therefore, machine learning approach might help to learn the algorithm to classify meaningful clusters. The distance metric on triclusters should be carefully chosen.
2. Scalability: the issue of scalability is quite challenging in the described technique, and is to be solved. In current state, the application can support only contexts with one long dimension, for instance, a context of 20 users, 20 tags, and 400000 resources which can be projected onto a user-tag plane. One possible way to address the scalability issue is to perform preliminary clustering of objects, attributes, and conditions separately, and then choose representatives from each cluster.
3. Extending the idea of a human-machine approach to other problems in FCA or data mining, such as exploring implications and association rules in order to find meaningful ones.

## 8 Conclusion

Visual analytics, as one of the flourishing domains of data analysis, can be useful in mining objects with similar attributes under similar conditions in a context of social network data. A special algorithm was developed for uniting such objects, attributes, and conditions in triclusters. The program framework under development is intended to graphically display the results of this algorithm and to empower an analyst to decide on the meaningfulness of clusters and tags or resources recommendations for objects.

**Acknowledgements** The author would like to thank his colleagues from Higher School of Economics Sergei Kuznetsov and Dmitry Ignatov for their well-timed advice and support during this work. He also expresses gratitude to Stanislav Klimenko from Moscow Institute of Physics and Technology for consulting in visual analytics.

## References

1. *Ganter, B., Wille, R.*: Formal concept analysis: Mathematical foundations. Springer, Berlin (1999)
2. *Lehmann F., Wille R.*: A triadic approach to formal concept analysis. Springer-Verlag, London (1995)
3. *Mirkin, B. G.*: Mathematical Classification and Clustering. Kluwer Academic Press, Dordrecht (1996)
4. *Ignatov, D. I., Kuznetsov, S. O., Poelmans, J., Zhukov, L. E.*: Can triconcepts become triclusters? International Journal of General Systems. 42, 572—593 (2013)
5. *Gnatyshak, D. V., Ignatov, D. I., Semenov, A., Poelmans, J.*: Analysing online social network data with biclustering and triclustering. In: Proceedings of the "Concept Discovery in Unstructured Data" conference, vol. 871, pp. 30—39. Katholieke Universiteit Leuven, Leuven (2012)
6. *Ignatov, D. I., Kuznetsov, S. O., Poelmans, J.*: Concept-Based Biclustering for Internet Advertisement. In: ICDM Workshops, pp. 123—130 (2012)
7. *Venjega, A. B., Gnatyshak, D. V., Ignatov, D. I., Konstantinov, A. V.*: Recommender system for perfumes and their tags based on triclustering. In: Proceedings of the "Intellectual data processing" conference, pp. 601—605. Torus Press, Moscow (in Russian) (2012)
8. *Ignatov, D. I., Poelmans, J., Zaharchuk, V.*: Recommender System Based on Algorithm of Bicluster Analysis RecBi. In: CEUR Workshop proceedings of the "Concept Discovery in Unstructured Data" conference, vol 757, pp. 122—126 (2011)
9. *Keim, D., Andrienko, G. et. al.*: Visual analytics: Definition, process, and challenges. In: Information Visualization, vol. 4950, pp. 154—175 (1999)
10. *Thomas, J., Cook, K.*: Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press, New York (2005)
11. *Kielman, J., Thomas, J.*: Special Issue: Foundations and Frontiers of Visual Analytics. In: Information Visualization, vol. 8, pp. 239—314 (2009)
12. *Kuznetsov, S. O.*: On Computing the Size of a Lattice and Related Decision Problems. Order, vol. 18, no. 4, pp. 313—321 (2001)

13. *Kuznetsov, S. O.*: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence. 49, 101—115 (2007)
14. *Gnatyshak, D. V., Ignatov, D. I., Kuznetsov, S. O.*: From Triadic FCA to Triclustering: Experimental Comparison of Some Triclustering Algorithms. In: CLA 2013 Proceedings, pp. 249—260. University of La Rochelle (2013)
15. *Kashnitsky, Y. S.*: Visual analytics for multidimensional data triclustering. Proceedings of MIPT, vol. 6, no. 2(22) (in Russian, to be published) (2014)

# Визуальная аналитика в задаче трикластеризации, основанной на анализе формальных понятий

Юрий Кашницкий

НИУ ВШЭ, Москва, Россия
ykashnitsky@hse.ru

**Аннотация** Трикластеризация — это способ обнаружения объектов со схожими свойствами в контексте из трех множеств сущностей. Например, в задаче анализа данных социальных сетей, такими множествами могут быть пользователи, их интересы и события, в которых они принимают участие. Трикластеризация здесь может помочь найти группы пользователей с похожими интересами и, делать им рекомендации событий на основе этих интересов. В статье описывается конкретный алгоритм трикластеризации и прототип программной платформы для визуального анализа полученных трикластеров.

**Ключевые слова:** визуальная аналитика, анализ формальных понятий, трикластеризация, анализ социальных сетей.