Unsupervised Co-Authorship Based Algorithm for Clustering of R&D Trends at Science and Technology Centers in Oil and Gas Industry

Fedor Krasnov[1], Mars Khasanov[2]

[1] Science & Technology Centre GazpromNeft,
75-79 liter D, Moika River emb., St Petersburg, 190000 Russia
krasnov.fv@gazprom-neft.ru
[2] Gazprom Neft PJSC, Pochtamtskaya ul. d. 3-5, St Petersburg 190000,Russia

Abstract. Planning of research and development trends in science and technology centers should be in line with the actual state of things. Such phenomena as organizational frigidity, research diversification and propensity for developing IT products are able to significantly impair any strategies and development trends.

However, feasibility of plans is an important attribute of development able to which significantly raise personnels motivation for achieving best results. This is why setting achievable goals is of such importance. There are never enough quantitative tools for appraisal of research and development activities. Formal paperwork reporting on R&D is not suitable for evaluation of researchers involvement and dedication.

Instead, small formats of research works such as presentations at scientific and technical conferences or scientific articles in peer-reviewed scientific publications require much more informal approach from researchers. Analysis of a science and technology centers performance based on its publication activity is a common practice. Many studies analyze text corpus of scientific articles and make conclusions on development trends. Text data noise levels are quite high; even most advanced analysis methods based on word embedding are able to produce accurate predictions only if analyzed are huge text volumes which are seldom available in case of small organizations. Small research organizations suffer the most from inaccurate planning of research activities.

Authors of this research propose to take advantages of articles (presentations) analysis based on co-authorship bipartite graph to extract research trends with the purpose of their further evaluation and planning.

Keywords: clustering, co-authorship graph, research activitys attributes, scientometrics, organizational hypotheses

1 Introduction

Todays focus on scientific approaches to managerial decisions becomes ever more vital. As data volumes grow analytic tools used by organizations management become less efficient. On the other hand often there is no sufficient data volume for sustainable work of advanced algorithms. On the forefront of this trend there is a problem of adaptation and developing new heuristics for solving such classic problems as clustering which are to be used for organizational purposes.

Data clustering based on a static model gained momentum as such algorithms when PAM [1], CLARANS [2], DBSCAN [3], CURE [4] and ROCK [5] had been discovered. However, lately a special focus is on the clustering algorithms based on dynamic model such as CHAMELEON [6]. The basic idea of the CHAMELEON [6] algorithm is in applying proximity metrics to a graph built on a set of clusterable data through the k nearest neighbor (KNN) method. Graph metrics prove more efficient for top-down data breaking in case of complex objects (Figure 1).

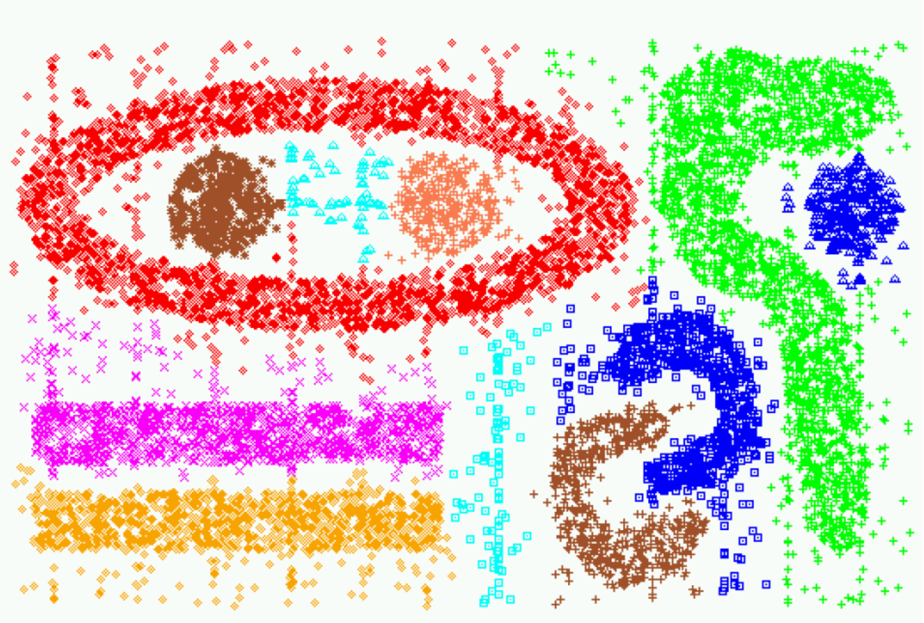

Fig. 1. Example of applying the CHAMELEON [6] algorithm for clustering of complex objects.

Algorithms diversity does not make less important the task of their efficiency evaluation. However, given a limited number of data and in order to improve managerial decisions the quality of clustering must to have not only mathematically substantiated but also reliable image components. In other words, it should be comprehensible at a glance and not requiring going deep into formulas. Such are the todays businesses needs.

2 Methodology of the research

From a formal point of view we have to solve the problem of unsupervised machine learning for co-authorship graph, attribute clusters to particular subjects and detect variations in clusters over the time.

Clustering of co-authorship graph can be achieved based on various nodes metrics:

- Degree centrality
- Betweenness centrality
- Closeness centrality
- Harmonic centrality
- Clustering

Let us examine the conceptual meaning of the Betweenness centrality metrics applied to the problem of clustering of co-authorship graph in an R&D organization. The Betweenness centrality metrics shows how important is a particular node for the graphs connectivity. Connections in a co-authorship graph reflect research collaboration. Co-authorship graphs are not always connected; usually they consist of several connected components of various sizes.

Connected components are natural clusters. Small connected components reflect primary initiatives researchers first articles. However the main connected component may contain up to 90% of a co authorship graphs nodes and call for a special approach to clustering.

To extract clusters from a main connected component of a co-authorship graph one may use the method of artificial removal of the nodes with the top-value Betweenness centrality metrics. As each of such nodes is removed a graph may break down into several disconnected components. The Figure 2 shows such separation model.

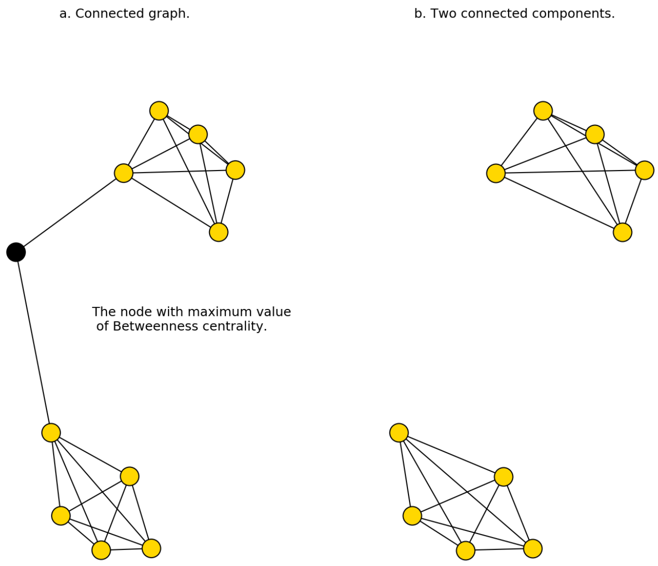
a. Connected graph.
b. Two connected components.
The node with maximum value
of Betweenness centrality.

Fig. 2. Graph separation model. . Initially connected graph. b. Same graph with the node with the top-value Betweenness centrality metrics removed looks like two connected components.

Each of the components resulting from such separation can be analyzed for subjects homogeneity based on articles texts of which each component is formed. Several iterations would result in a set of clusters.

The method proposed by the authors is a heuristic one and requires examination by a particular formal criterion. Conventional criteria for the purposes of clustering are proximity metrics for a cluster components and distances between components in separate clusters.

Convergence of the authors method is ensured through searching a minimum of functional errors in determining k clusters with 1.

$$\frac{WSS}{BSS} - > min \tag{1}$$

Where WSS_{c_i} within-cluster variation for cluster C_i, m_i - centroid of C_i and $i \in [1..k]$ (2). The total WSS measures the compactness of the clustering and we want it to be as small as possible.

$$WSS = \sum_{i}^{k} \sum_{x \in C_i} \left(x - m_i\right)^2 \tag{2}$$

And BSS - weighted inter-cluster separation, measured by the between cluster sum of squares (3).

$$BSS = \sum_{j}^{k} \sum_{i}^{k} |C_i| \left(m_j - m_i\right)^2 \tag{3}$$

Where $|C_i|$ - is a cluster size.

Interdisciplinary researches lead to the situation where articles may fall into several subject categories, thus the resulting clusters would be intersecting and non-exclusive.

3 Results

The Gazpromneft R&D Center‘s publication activity has been chosen as a research subject. The data has been obtained from the OnePetro open online library of the international Society of Petroleum Engineers (SPE). Upon cleansing 172 articles have been singled out.

Let us base our prediction on a co-authorship graph. For this purpose we build a co-authorship bipartite graph [7] with the nodes: author (479) and article (171). Authors have technical competences while articles have such attributes as title, year of publication and key words.

The resulting co-authorship graph has 26 connected components of which the strongest one has 556 nodes while the others have maximum eight nodes. Connected components with up to eight nodes represent the researchers first articles.

Let us examine the strongest connected component (556 nodes). We extract a subgraph from the main co-authorship graph based on the nodes contained in the strongest connected component. The resulting subgraph is shown on the Figure 3 .

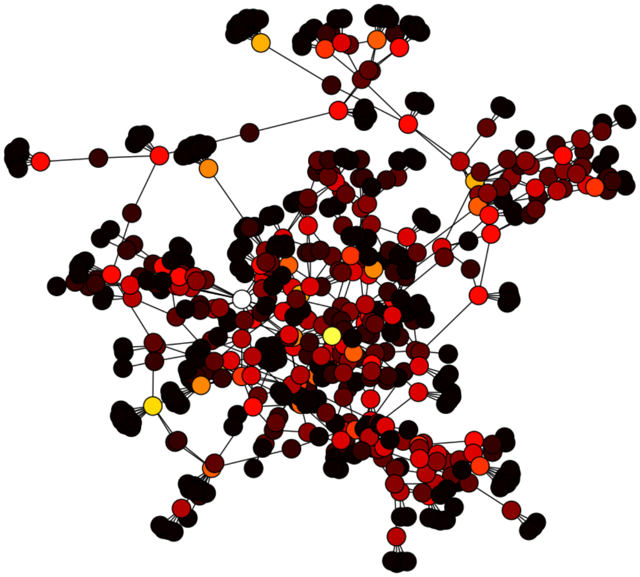

Fig. 3. Subgraph of the strongest connected component of the co-authorship graph of Gazpromneft R&D center

Let us compute the Betweenness centrality metrics for the resulting subgraph. The obtained Betweenness centrality values are shown on the Figure 4. Zero values for the Betweenness centrality are not shown.

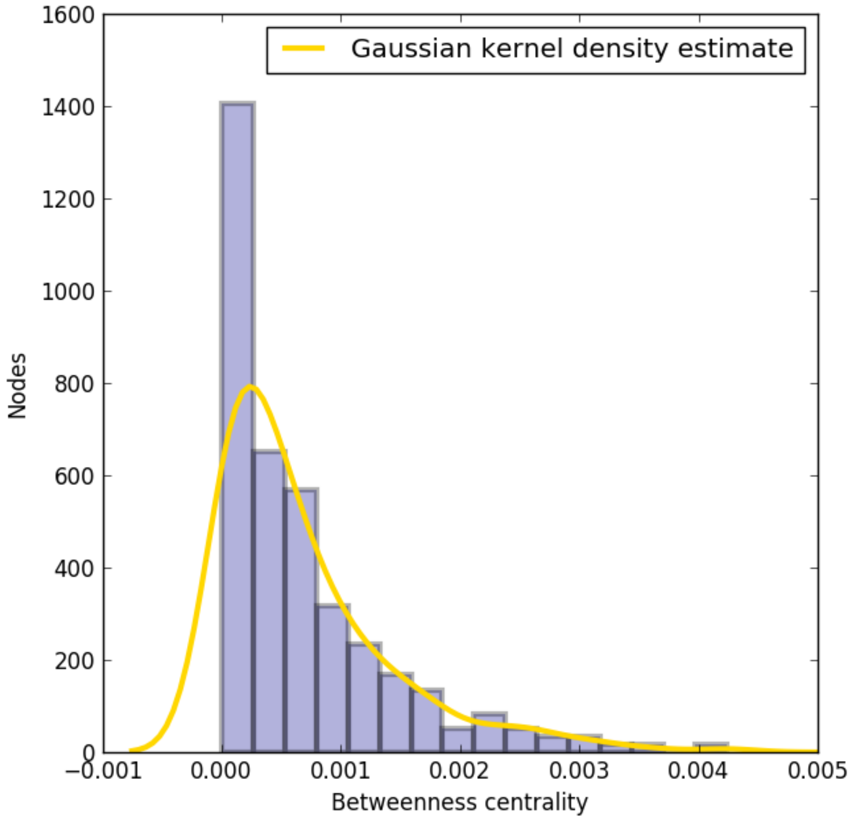

Fig. 4. Histogram of Betweenness centrality values for the subgraph of the strongest connected component of the co-authorship graph of Gazpromneft R&D center

As we can see on the Figure 4 the values of the Betweenness centrality metrics in the third quartile belong to only 23 nodes which represent less than 5% of the total number of nodes.

Let us apply the algorithm of artificial removal of the nodes with the highest value of the Betweenness centrality metrics. The Figure 5 shows correlation between the connected components number and the number of artificially removed nodes.

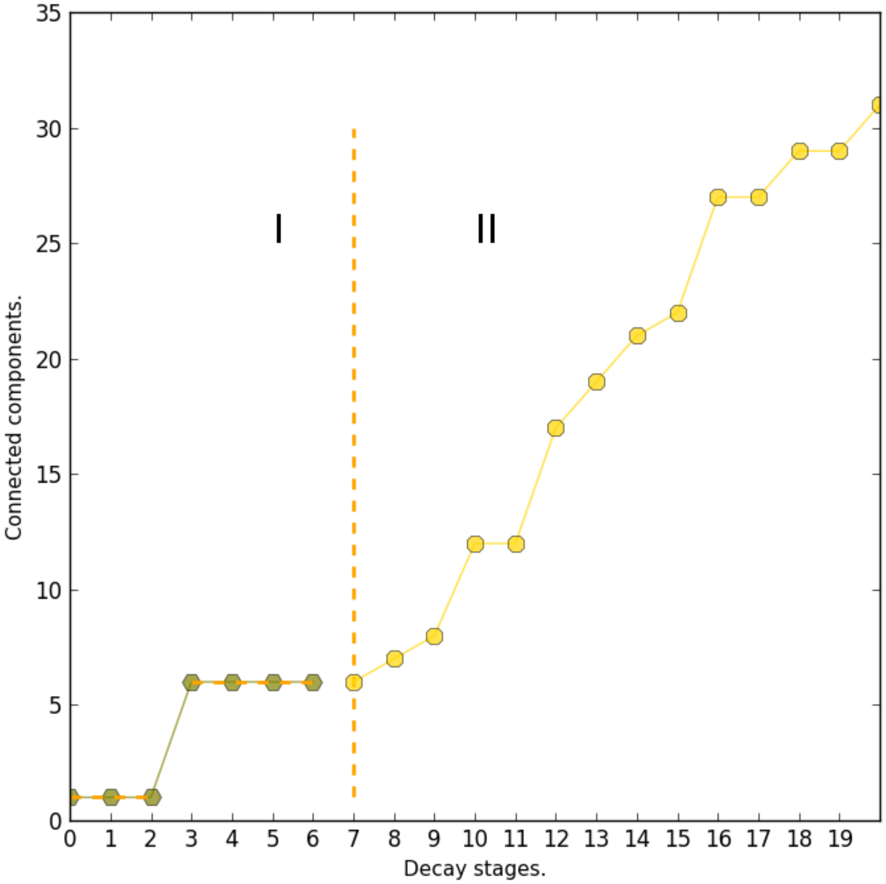

Fig. 5. Correlation between the connected components number and the number of artificially removed nodes.

As the nodes get removed the graph can behave in two following modes:

1. Connectivity constraint (Mode I)
2. Exponential decay (Mode II)

Mode I is characterized by the graphs retaining its connectivity as the nodes with high values of the Betweenness centrality metrics get removed. It means that the removed nodes are not the only connections between clusters.

Mode II is characterized by following the exponential model of a graphs decay when each removed node causes exponential growth in emergence of new connected components.

Let us have a closer look at the second half of the Mode I of the algorithm when the graph has broken down into six connected components. These components sizes are 511, 34, 1, 1, 1, 1. Among them the component with 34 nodes shown on the Figure 6 represents the most pronounced direction of research into Subject 1.

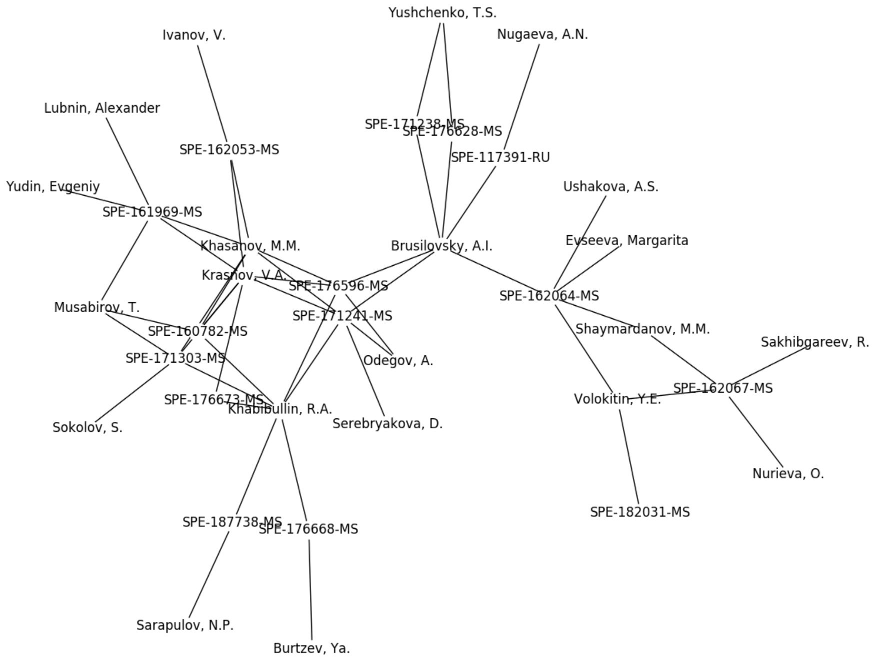

Fig. 6. The cluster of researchers into Subject 1 extracted through the method of removal of the nodes with the highest values of the Betweenness centrality metrics.

We have examined extraction of one cluster in detail. The complete algorithm of clusters extraction would consist of the following steps:

1. Building a co-authorship bipartite graph: G
2. Finding the Betweenness centrality metrics for the G graph
3. Finding a node with BC_{max} metrics (Betweenness centrality)
4. Removing the BC_{max} node (Betweenness centrality) from the G graph
5. Deriving a list of connected components of the G graph
6. Computing a quality metrics WSS and BSS of the retrieved clusters
7. Further the algorithm is iterated for each connected component
8. Algorithm is completed when all connected components represent clusters of acceptable quality.

For the selected co-authorship graph 16 clusters have been extracted. To compute values and W based on the articles texts we have applied the Vector Space Model (VSM). Each article is represented as a vector with the BM25 [8] metrics values for each word. Articles are considered as BOW ("bag of words"). For measuring distances between the articles VSM we have applied cosine measure. The Figure 7 shows the clusters separability matrix.

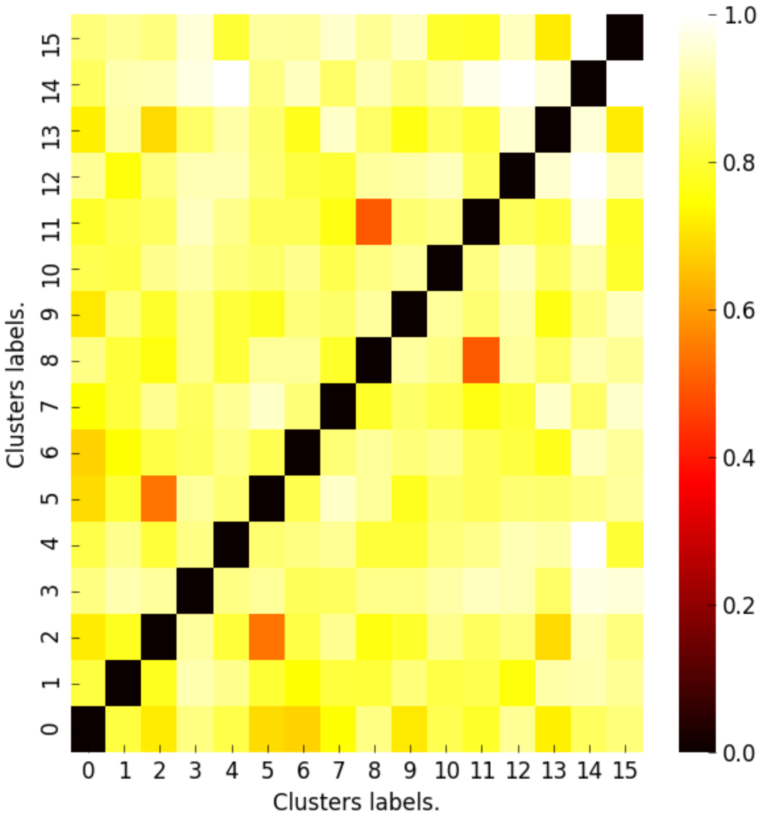

Fig. 7. Clusters separability matrix. Clusters numbers are on the axes. BSS function values are in the cells.

For the purposes of comparison of the resulting articles clustering we have performed clustering with the KMeans algorithm which yielded similar results (Figure 8).

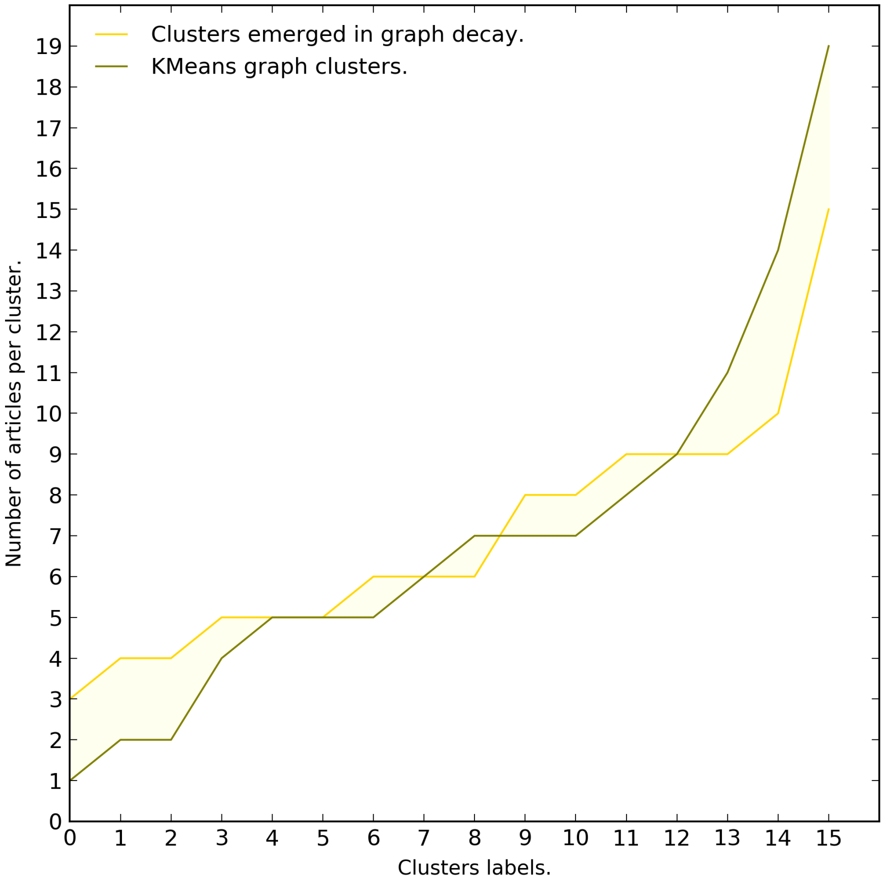

Fig. 8. Comparison of the clustering algorithm proposed in this article with the KMeans algorithm.

The articles corpus has been broken down into clusters using the KMeans algorithm. The resulting clusters allowed arranging authors into groups.

4 Conclusion

The authors have proposed a method of extraction of research trends based on the co-authorship graph. Concept-wise the method belongs to the top-down clustering algorithms. The Betweenness centrality metrics has been chosen as a criterion for extracting clusters.

The metrics of cluster components proximity and the metrics of distances between separate clusters based on the subjects of articles in the co-authorship graph have been applied as a clusters quality criterion. This method resulted in an aggregate vision of organizations research trend based on the open data on its researchers publication activities.

The authors method of extraction of research trends based on the co-authorship graph has been tested at the Gazpromneft R&D Center. As a result 16 clusters indicative of the organization activity have been extracted. The following attributes of the authors method of extraction of research trends based on the co authorship graph are significant:

- Recursive algorithm allows working with graphs of various orders.
- Greedy algorithm for clusters quality evaluation allows correcting optimization at any step.
- Applying a co-authorship bipartite graph allows analyzing various projections.
- Working with the data in public domain gives ample opportunities for application in business intelligence.

The novelty of the method of extraction of research trends based on the co-authorship graph proposed by the authors is in applying the co-authorship bipartite graph and in the dynamic model of clustering using structural metrics for the co-authorship graph and proximity metrics for research articles texts.

References

1. Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.
2. Raymond T. Ng and Jiawei Han. Clarans: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering*, 14(5):1003–1016, 2002.
3. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
4. Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: an efficient clustering algorithm for large databases. In *ACM Sigmod Record*, volume 27, pages 73–84. ACM, 1998.
5. Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock: A robust clustering algorithm for categorical attributes. *Information systems*, 25(5):345–366, 2000.
6. George Karypis, Eui-Hong Han, and Vipin Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.
7. Fedor Krasnov. Analysis of methods of construction of the graph of co-authorship: an approach based on bipartite graph. *International Journal of Open Information Technologies*, 6(2):31–37, 2018.
8. Yuanhua Lv and ChengXiang Zhai. Adaptive term frequency normalization for bm25. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1985–1988. ACM, 2011.