Detection of the thematic groups in scientific publications

Pavel Kozlov[a], Andrey Mokhov[a] and Vladimir Tolcheev[a]

a National Research University "Moscow Power Engineering Institute", Krasnokazarmennaya 17, Moscow, 11250, Russian Federation

Abstract

The paper identifies thematic groups (clusters) in scientific publications of members of small research teams. The sample is formed from the articles contained in the Russian digital library eLibrary.ru. Text documents are pre-processed and their mathematical description is given. Using the algorithms of exploration analysis we detect the structure of the initial sample. Then we define the main topics of research team and analyze the results obtained by different cluster methods.

Keywords 1

Hierarchical cluster analysis, clustering, K-means and k-means++ method, visualization, cosine proximity measure.

1. Introduction

In our article we examine the thematic proximity of scientific publications of members of small research teams (for example, department or laboratory). Most often, such teams conduct research on fairly close overlapping topics, so the detection of clusters has to be performed in conditions of small differences in the terminology of documents. To get reliable results we use various methods of exploratory analysis and define thematic groups of publications (clusters).

The extraction of clusters allows us to detail the areas of specialization of the research team. It is useful for customers interested in conducting research on specific topics potential customers (primarily partners from industry) and students, who want to get knowledge on subjects that are in demand in practice.

In our work, we identify thematic groups in scientific publications of specialists of the Department of Control and Intelligent Technologies (CIT) of the National research University "Moscow power engineering institute" (NRU "MPEI"). A preliminary review of the training courses of the CIT allows us to assume that department is specialized in the following areas: "Theory of automatic control, simulation, identification, optimization", "Data analysis, information security, information and analytical systems", "Microprocessor technology and SCADA systems".

2. Problem statement and preliminaries

We create sample by selecting Department's publications indexed in the Russian digital library eLibrary.ru (https://www.elibrary.ru/) in the period 1991-2019. This library provides free access to bibliographic descriptions of articles, including titles, annotations, and keywords. In our work, we analyze primarily Russian-language publications and the sample does not include English-language articles in foreign journals, any patents, certificates of registration of computer programs, acts of implementation, research reports, dissertations. In our opinion, this does not have a significant impact

Russian Advances in Fuzzy Systems and Soft Computing: selected contributions to the 8-th International Conference on Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT-2020), June 29 – July 1, 2020, Smolensk, Russia
EMAIL: kozlov.pavel.andreevih@yandex.ru; asmokhov@mail.ru; tolcheevvo@mail.ru;
ORCID: 0000-0002-1979-6411 (A. 2);

on the quality of the data, since almost all English-language materials most often have in eLibrary.ru Russian-language duplicates.

For the Department of Control and Intelligent Technologies the sample {X} consists of 351 publications (n =351) made by 13 scientists of the Department (N=13). {X} does not include articles of specialists who (on September 2019 - the time of sampling) had fewer than 5 publications indexed in eLibrary.ru. Thus we exclude from the set {X} scientific works of postgraduates and students who are not employees of the Department as well as specialists who are poorly involved in research. Unfortunately, in some cases, we face incomplete data due to rather slow indexing of articles in eLibrary.ru.

Let's make a formal statement of the problem: there are many (Russian-language) publications of the Department - {X}, contained in the eLibrary.ru, and a set of topics (research areas) that these publications correspond to $\{Q_1,..., Q_k, ... Q_K\}$. Moreover, the number of topics (and their names) is unknown in advance [1,2,3]. In our article we consider non-overlapping hard clustering, in which each document belongs to only one cluster.

The challenge before us is to determine the number of clusters and their names (direction of research). In addition we need to assign each document to a specific cluster.

We can solve this problem in the following ways:

- Perform clustering of the original sample and define groups of terminologically similar publications (create clusters containing publications);
- Conduct clustering of specialists of the Department, for example, on the basis of co-authorship (get clusters consisting of scientists who have joint publications);
- Carry out clustering of terms and build "clouds" of strongly related terms.

In our article, the "bag of words" and a vector representation are used to describe a text document. For this stop words and rare terms (no more than twice) are removed from dataset {X} and lemmatization is performed using the Python library (pymorphy2), which returns all words to their initial form. Each document X_m from sample {X} is described by a vector that includes the frequency of terms calculated from bibliographic descriptions. The dimension of all vectors is the same and is equal to the L-number of informative terms in the sample (in our research L = 1500).

$$X_m = \begin{bmatrix} x_1{}^{(m)} \\ x_l{}^{(m)} \\ x_L{}^{(m)} \end{bmatrix}, \quad (m = 1,..,n; l = 1,..,L) \tag{1}$$

For further research we use the following matrix descriptions of the initial sample:

1. The matrix of “term – term”:

$$B = \begin{pmatrix} b_{11} & .. & b_{1H} \\ .. & b_{ph} & .. \\ b_{H1} & .. & b_{HH} \end{pmatrix}, \; p,h = 1..H \tag{2}$$

Where b_{ph} – the frequency of joint occurrence of the p-th and h-th terms (p≠h) and b_{pp} –the frequency of occurrence of the p-th term when p=h, H – the number of keywords in the sample (note that here only keywords (a part of the bibliographic description) are used to describe documents).

2. The matrix of “author – author”:

$$\mathrm{A} = \begin{pmatrix} a_{11} & .. & a_{1N} \\ .. & a_{ij} & .. \\ a_{N1} & .. & a_{NN} \end{pmatrix}, \tag{3}$$

Where a_{ij} – the number of authors ' joint publications (i≠j), a_{ii} – the number of author's publications (i=j), N – the number of co-authors $(i,j = 1,\dots,N)$.

3. The matrix of “term-document”:

$$X = \begin{pmatrix} x_{11} & .. & x_{1n} \\ .. & x_{lm} & .. \\ x_{L1} & .. & x_{Ln} \end{pmatrix}, \; (\mathrm{m} = 1,..,\mathrm{n}; \mathrm{l} = 1,..,\mathrm{L}) \tag{4}$$

Where x_{tm} is calculated by weighing TF (term frequency), i.e. the term weight is calculated as the frequency of occurrence of an informative word [1].

$$x_{tm} = \frac{x_{tm}}{\sum_{l=1}^{L} x_{lm}} \quad (5)$$

3. Identification of thematic groups and analysis of results

We will perform an exploratory analysis of the initial sample based on the authorship-co-authorship graph. Usually in the literature, the calculation of joint publications is used to identify thematically related groups of scientists [4]. In our study, as a mathematical description of the sample, we use the "author-author" matrix, which shows the number of joint articles. Visualization of the resulting graph is obtained using the Gephi program [5]. Setting the number of author's publications as the vertex size, and the number of shared articles as edges, we get the graph (

Figure **1**):

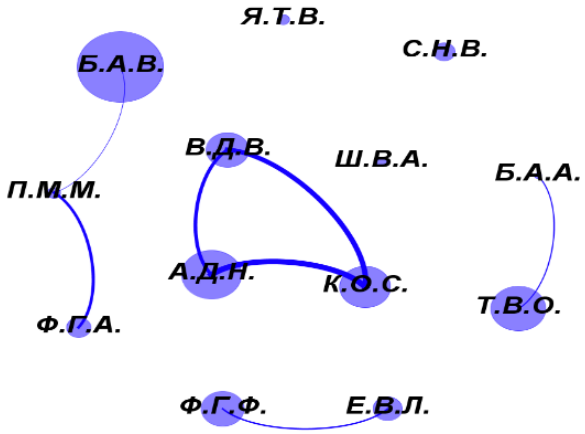

Figure 1: Graph of authorship and co-authorship

In Figure 1, the point names are the full names of the authors. According to it, the Department can be divided into 4 groups of specialists who have joint publications. These groups fall into two categories: “Joint research” or “Scientific leader – the disciples.”

Let's interpret the results on figure 1 on base of analysis of the most common words in each cluster. In Table 1 we present ten most important (high-frequency) terms for each cluster and give the number of publications in each cluster.

Table 1

Most frequent words of clusters

Cluster #1	Cluster #2	Cluster #3	Cluster #4
INFORMATION	SYSTEM	CONTROL	SYSTEM
ANALYSIS	CONTROL	SYSTEM	TASK
METHOD	MODEL	INFORMATIONAL	CONTROL
PROCESSING	FUZZY	OBJECT	METHOD
SEARCH	MODAL	MODEL	ANALYSIS
INFORMATIVE	OBJECT	FACTOR	DETECTION
SEARCH	DYNAMIC	DECISION	NEURONET
TYPE	METHOD	PROCESS	ALGORITHM
CLASSIFICATION	ADAPTIVE	PROCESSING	OBJECT
CLASSIFIER	IDENTIFICATION	COLLECTION	MODEL
Number Of Articles	Number Of Articles	Number Of Articles	Number Of Articles
42	149	98	62

Note that the clusters differ quite significantly in the number of publications. The smallest cluster (Б.А.А. и Т.В.О.) contains 42 articles, while the largest cluster (А.Д.Н., В.Д.В., К.О.С.) includes almost 4 times as many publications. Despite the different size of clusters they are fairly well interpreted and correspond to the following topics:

1 cluster: “search, analysis, processing and classification of text information”;

2 cluster: “(fuzzy) control of dynamic systems and their identification”;

3 cluster: “information systems and decision-making systems”;

4 cluster: “neural networks in control and data processing”.

The cluster split obtained in figure 1 generally corresponds to expert estimates about Department’s specialization. However, based on the authorship-co-authorship graph, it is difficult to make a conclusion about the number of clusters that combine thematically similar publications. It is impossible to exclude cases when several groups of specialists identified on the figure 1 conduct independent research in the same scientific direction.

For this reason, we use hierarchical and non-hierarchical (flat) clustering methods in further research. These methods (in contrast to more complex procedures, such as latent semantic analysis and its modifications) are very good at separating small samples.

First we will apply hierarchical cluster analysis and combine scientists not based on co-authorship, but by building their terminological profiles.

A profile is a vector whose components are weights calculated as the frequency of occurrence of terms in the author's publications [6,7].

$$Y_j = \begin{bmatrix} y_1^{(j)} \\ \dots \\ y_L^{(j)} \end{bmatrix}, (j = 1, .., N) \quad (6)$$

The built profiles are then combined using a dendrogram. The Python Sklearn library is used for this purpose. The proximity between profiles is calculated using the cosine measure [1]:

$$\cos\alpha = \frac{(Y_j, Y_i)}{|Y_j| * |Y_i|} = \frac{\sum_{l=1}^{L} y_l^{(j)} * y_l^{(i)}}{\sqrt[2]{\sum_{l=1}^{L} (y_l^{(j)})^2} * \sqrt[2]{\sum_{l=1}^{L} (y_l^{(i)})^2}} \quad (7)$$

The above formula uses the following notation: $y_l^{(j)}$ - frequency of the l-th word in the profile of the j-th specialist, $y_l^{(i)}$ – frequency of the l-th word in the profile of the i-th specialist ($i, j = 1, .., N$).

In our research we analyzed various ways to combine clusters, which gave quite similar results. On all dendrograms, from 3 to 5 groups are consistently distinguished at various levels of detail.

Figure **2** shows the result of hierarchical clustering for the cosine measure and the complete linkage method. The groups shown in figure 1 (authorship-co-authorship graph) and figure 2 (dendrogram) agree well. In figure 2 single clusters (“С.Н.В.”, “Я.Т.В.”, “Ш.В.А.”) join larger clusters. At the same time, the cluster analysis revealed a terminological similarity between the publications of authors who do not have common articles (“Я.Т.В.” и “Ш.В.А.”). But they both are working in the field of control theory and automation. At the same time, it should be noted that “Ш.В.А.” conducts scientific researches in different directions. His interests include not only the theory of automatic control, but also reliability, energy saving, artificial neural networks and cyber-physical systems. For this reason “Ш.В.А.” is located on the border of several thematic clusters and (depending on the parameters of hierarchical cluster analysis) is attracted to different clusters.

The cosine measure allowed us to obtain significantly more interpreted results than using the euclidean metric and its modifications. With varying ways of combining clusters (single linkage, complete linkage, unweighted pair group average, weighted pair group average, weighted centroid pair group, Ward method) was observed moving “Я.Т.В.” and “Ш.В.А.” between clusters and isolation of “С.Н.В.” in a separate stand-alone cluster.

The analysis of the results presented in figures 1 and 2 shows a fairly simple structure of clusters without strong inter-cluster connections, which allows us to conclude that there are no interdisciplinary studies and large projects that involve the majority of specialists of the Department.

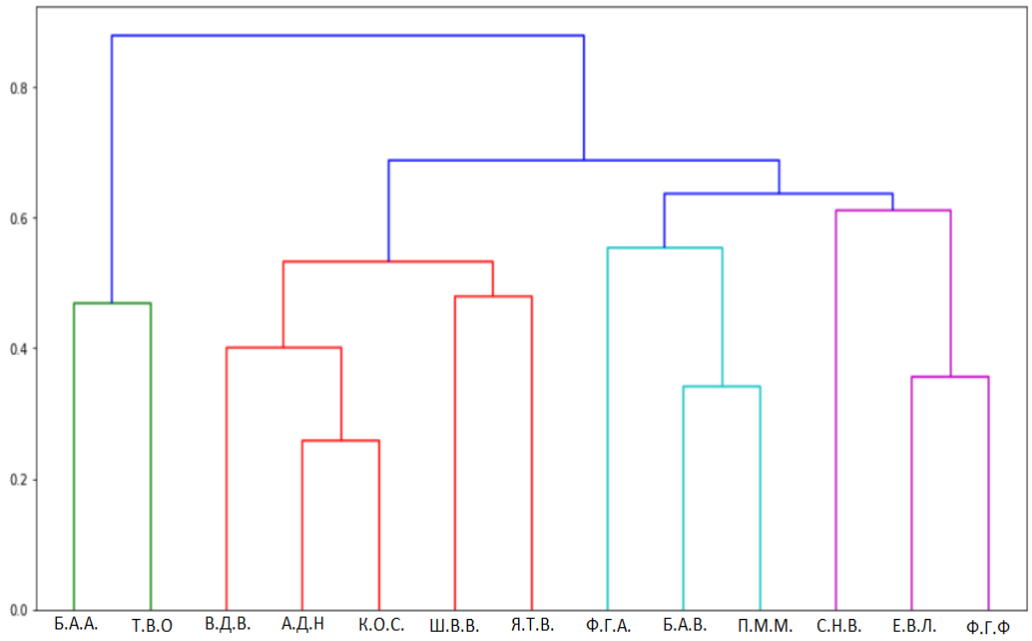

Figure 2: Dendrogram for the cosine measure and complete linkage method

Taking into account the specifics of the clustering problem, in which we do not know the only correct division of publications (authors, terms) into groups, it is necessary to conduct additional researches in order to identify a stable division into groups. In our paper for this purpose we apply the well-known k-means method. This method is based on minimizing the sum of squares of distances within clusters [1,8]:

$$V = \sum_{k=1}^{K} \sum_{X \in Q_k} (X - \mu_k)^2 \qquad (8)$$

X is the frequency vector of document terms that belong to the k-th cluster, and μ_k is the centroid of the k-th cluster.

This method (as well as hierarchical cluster analysis) has two configurable parameters - the measure of proximity and k - the number of clusters that are split. However, there is another important problem that must be solved before the research begins - how to choose the initial position of the centroids.

To select the initial centroids we use “k-means++” method. In this variant of the k-means method, the first cluster centroid is selected randomly from data points, and then each next centroid is selected depending on the value of the square of the distance to the nearest (already selected) centroid. This approach allows to select the initial centroids more effectively compared to their random selection and more quickly determine the parameter k, which minimizes the sum of intra-cluster distances.

Let us refine the parameter k (number of clusters) by calculating the silhouette coefficient for different numbers of clusters. The value of the silhouette coefficient for a cluster element is determined using the formula [9]:

$$S_i = \frac{C_i - D_i}{\max(C_i, D_i)}, (i = 1, .., n) \qquad (9)$$

Where C_i is the average distance from the i-th object to objects from the same cluster, and D_i is the average distance from the i – th object to objects from the nearest other cluster.

The silhouette of the sample is the average value of the silhouette coefficients for objects in this sample. It shows how the average distance to objects in one cluster differs from the average distance to objects in other clusters. This value is in the range [-1.1]. Values close to “-1” correspond to

scattered clustering results. If the values are set to “0”, then the clusters intersect. Values close to “1” correspond to clearly defined clusters.

We calculated the silhouettes coefficients in the range from 1 to 15 clusters. The best value is obtained for k = 7 (

Figure **3**). With this number of clusters is possible to avoid the creation of very large and very small groups of documents. However, the values of the silhouette coefficient indicate that there are intersections between the formed clusters.

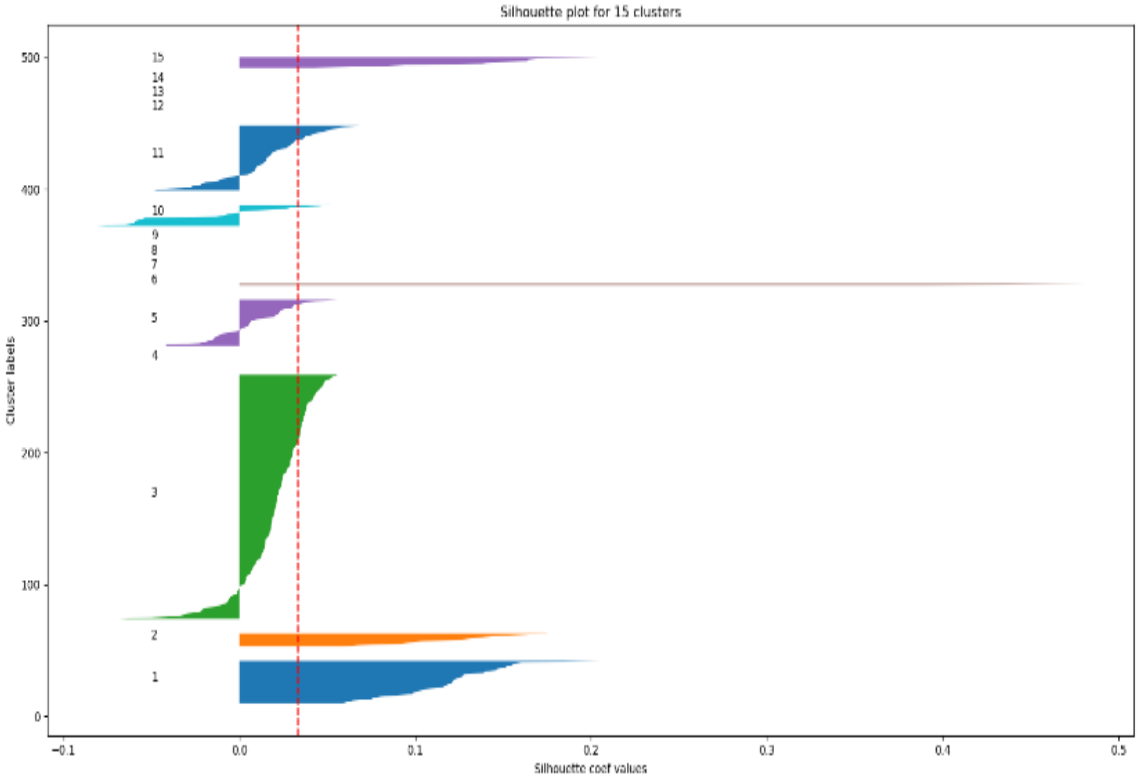

Figure 3: Silhouette coefficients for 7 clusters

The analysis of the most frequent words allows you to give the following names to clusters obtained using the k-means method (for k=7):

1 cluster: “Information systems in the field of Economics and financial monitoring”;

2 cluster: “Neural networks in control, simulation and reliability problems”;

3 cluster: “Analysis and processing of text information”;

4 cluster: “Automatic control systems”;

5 cluster: “Classification and diagnostics methods”;

6 cluster: “Fuzzy systems”;

7 cluster: “System identification”.

In this split, several overlapping groups are obtained that combine publications from the same subject area. Clusters # 3 and # 5 may seem almost identical, which contain several common high-rating terms (for example, "classification"). To explain this result, it is necessary to conduct an expert study of the publications, which revealed the presence of several scientific groups conducting independent research in the field of Data Analysis, in particular classification.

The appearance of clusters #3 and # 5 reflects not only the terminological differences in the articles, but also the different nature of the data used (cluster # 3 processes mainly text documents, while cluster # 5 processes factual (in particular medical) information). You can assume that when the sample size increases, both clusters are merged. This also confirms the observation that most authors' articles are distributed in different clusters (on average, they fall into three groups). In addition, we note that the creation of clusters (and their subject matter) is significantly influenced by scientists who have a significant number of publications or publish articles on a single (distinct) topic.

Clustering methods also allows to analyze the terminological differences of publications and identify groups of words that most often occur together. The term-term matrix is used as a mathematical model for describing the sample. To visualize the dictionary of terms in this work, we used the Gephi program, which previously analyzed the authorship-co-authorship graph. The main problem of graphs obtained using the Gephi program is that they are heavily overloaded with terms and relationships, which makes it difficult to identify and analyze existing patterns [11]. To simplify the graph, we took into account only high-frequency words and introduced a limit on the maximum number of edges originating from a single vertex.

As result of visualization we get three groups of terms (

Figure **4**): the blue group combines the topics “fuzzy systems and regulators + system identification and modal control”. The green area describes the terms most commonly found in the clusters “analysis and processing of information + classification methods + information systems in the field of Economics and financial monitoring”. As for the red section of the graph, it combines the terms “neural networks in management, simulation, reliability + control of systems and objects”. In fact, the Gephi program has divided the terminology space of the Department’s publications into three large groups: “Control and neural networks” (red cluster), “Data Analysis and information systems” (green cluster), “Fuzzy models and System identification” (blue cluster). At the same time, almost no publications were found on the subject of “Microprocessor technology and SCADA systems” (in this subject, as indicated earlier, the Department of Control and Intelligent Technologies reads specialized training courses).

Thus different clustering methods applied to the original sample don’t provide completely identical results. The lack of an unambiguous answer when identifying the structure of text arrays is a well-known situation [9], which is typical for exploratory data analysis. However, in our approach, changing the number of clusters (from 3 clusters to 7) allows us to study the original sample with the required level of detail, depending on the researcher's preferences. It should also be noted despite some differences, cluster partitioning obtained in different ways (matrix of “term – term”, matrix of “author – author”, matrix of “term-document”) does not contradict and complement each other.

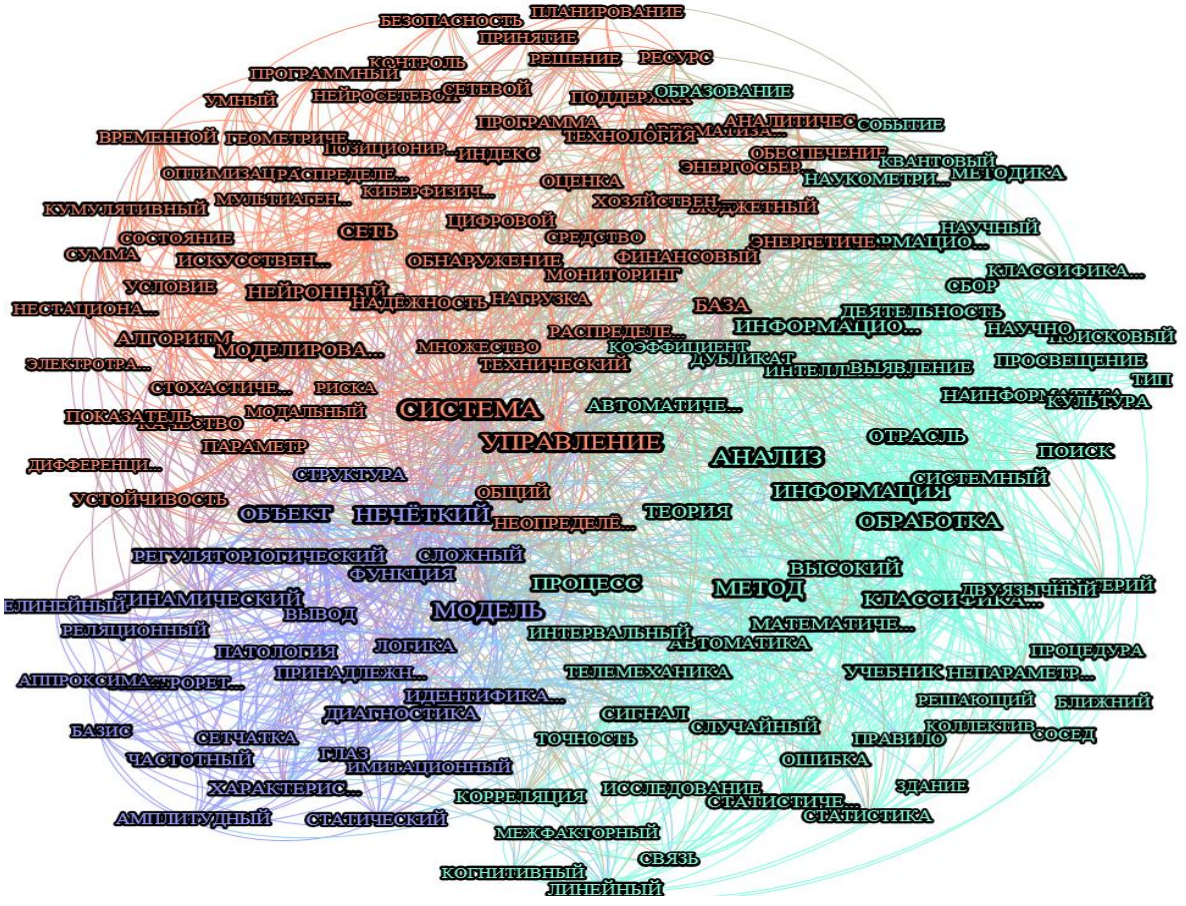

Figure 4.Term clustering using Gephy

4. Conclusions

Our research allows us to obtain well-interpreted results and extract thematic groups of publications from a set of scientific papers of a small research team (the results are illustrated in relation to the Department of Control and Intelligent Technologies of Moscow power engineering institute). For this we use different clustering approaches and investigate problem on different levels of detail. We give the names of the resulting clusters and compare them with the expert division into thematic groups. Clustering methods and expert evaluations are almost identical. Important to note that comparison of the Department's lecture courses and research specialization also showed high consistency. The same we can say about topics of bachelor's (and master's) works and research directions by postgraduates.

The results obtained by us are currently being used to develop a recommendation system that will allow to search publications in the Russian digital library eLibrary.ru. and identify articles that best meet the information needs of specialists of Department of Control and Intelligent Technologies, correspond to their terminology profiles and scientific interests.

In further research, it is planned to analyze the degree of changes in topics over time, as well as apply alternative approaches (in particular, latent semantic analysis) to obtain clusters that combine authors and topics.

5. References

[1] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008, ISBN: 1139472100, 9781139472104.

[2] N. A. Astrakhantsev, D. G. Fedorenko, D. Yu. Turdakov, Methods for automatic term recognition in domain-specific text collections: A survey. Programming and Computer Software, 41, 6 (2015) 336-349.

[3] K. W. Boyack, R. Klavans, Accurately identifying topics using text: Mapping PubMed. In R. Costas, T. Franssen, & A. Yegros-Yegros (Eds.), Proceedings of the 23rd International Conference on Science and Technology Indicators. Leiden, the Netherlands, 2018, 107-115.

[4] L. Subelj, N J van Eck, L. Waltman,Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods. PLoS ONE , 11(4) (2016). doi:10.1371/journal.pone.0154404.

[5] The Open Graph Viz Platform (https://gephi.org/).

[6] A. Mokhov, V. Tolcheev, Automated Construction and Analysis of Scientific and Educational Profiles of the University Department, V International Conference on Information Technologies in Engineering Education (Inforino), Moscow, Russia, 2020, 1-4, doi: 10.1109/Inforino48376.2020.9111842.

[7] P. A. Kozlov, A. S. Mokhov, V. O. Tolcheev, Viyavlenie tematik kafedralnikh publikaciy sredstvami Text Mining, Information technologies and informational security in science, technics and education “INFOTEX - 2019”. SevGU 2019, 124-128.

[8] A. J. Patton, B. M. Weller, Testing for Unobserved Heterogeneity via k-means Clustering, 2019. Papers 1907.07582, arXiv.org.

[9] P. Flach, Machine Learning: The Art and Science of Algorithms that Make Sense of Data. University Press, 2012. ISBN 10:1107422221.

[10] S. Fortunato, Community detection in graphs. Physics Reports, 486, 3-5 (2011) 75-174.

[11] J. Glaser, A. Scharnhorst, W. Glanzel, Same data—different results? Towards a comparative approach to the identification of thematic structures in science. Scientometrics, 111, 2 (2017) 981-998.