*Пропонується метод кластеризації публікацій науковців за науковими напрямами. В рамках даного методу запропоновано два способи знаходження відстані між публікаціями. Перший спосіб використовує довжину маршруту у графі цитування між публікаціями. Другий спосіб враховує розрахунок подібності між анотаціями публікацій на основі методу локально-чутливого хешування. Також пропонується метод ідентифікації напрямів досліджень науковців, який базується на результатах кластеризації наукових публікацій*

*Ключові слова: кластеризація, напрям наукових досліджень, граф цитувань, локально-чутливе хешування*

*Предлагается метод кластеризации публикаций ученых по научным направлениям. В рамках данного метода предложено два способа нахождения расстояния между публикациями. Первый способ использует длину маршрута в графе цитирования между публикациями. Второй способ учитывает расчет подобия между аннотациями публикаций на основе метода локально-чувствительного хеширования. Также предложен метод идентификации направлений исследований ученых, который базируется на результатах кластеризации научных публикаций*

*Ключевые слова: кластеризация, направление научных исследований, граф цитирования, локально-чувствительное хеширование*

UDC 005.8

DOI: 10.15587/1729-4061.2017.112323

# A METHOD FOR THE IDENTIFICATION OF SCIENTISTS' RESEARCH AREAS BASED ON A CLUSTER ANALYSIS OF SCIENTIFIC PUBLICATIONS

**A. Biloshchytskyi**
Doctor of Technical Sciences, Professor
Department of Network and Internet Technologies
Taras Shevchenko National University of Kyiv
Volodymyrska str., 60, Kyiv, Ukraine, 01033
E-mail: bao1978@gmail.com

**A. Kuchansky**
PhD, Associate Professor*
E-mail: kuczanski@gmail.com

**Yu. Andrashko**
Lecturer***
E-mail: yurii.andrashko@uzhnu.edu.ua

**S. Biloshchytska**
PhD, Associate Professor
Department of Information Technology Designing and Applied Mathematics**
E-mail: bsvetlana2007@ukr.net

**O. Kuzka**
PhD, Associate Professor***
E-mail: oleksandr.kuzka@uzhnu.edu.ua

**Ye. Shabala**
PhD, Associate Professor*
E-mail: wild_miledi@ukr.net

**T. Lyashchenko**
Senior Lecturer
Department of Information Technologies**
E-mail: liazschenko@ukr.net

*Department of Cybersecurity and Computer Engineering**
**Kyiv National University of Construction and Architecture
Povitroflotskyi ave., 31, Kyiv, Ukraine, 03037
***Department of System Analysis and Optimization Theory
Uzhhorod National University
Narodna sq., 3, Uzhhorod, Ukraine, 88000

## 1. Introduction

Economic growth and prosperity of any country depends largely on the development of science, technology, efficiency of the productive forces of society, etc. Certainly, the scientific-technical progress is impossible without ensuring the quality and efficiency of scientific research. Therefore, the establishment of criteria for the evaluation of research activities, emphasis on the analysis of scientific areas tackled by researchers, are important tasks for scientific and educational institutions, companies, which are engaged in the creation of scientifically-intensive technologies, and the state in general.

In order to effectively evaluate the quality of research work within the University, region, or state, it is necessary to first determine the relationships between all scientific publications by analyzing citations between these publications. Identified links between scientific publications open up possibilities for clustering these publications according to research areas. Every scientific direction would in this case include a certain number of authors of publications that belong to a particular direction or cluster. Thus, it is possible to match a particular author with a certain number of research areas, that is, to perform identification of the authors' research directions. The potential of a scientific area will be measured by assessing the results of the research activities of scientists related to this direction. Essentially, the clusters or areas of scientific research can be considered as dynamic objects that have the history of development and predicted potential.

The tasks of designing a method for the identification of areas of research of scientists, as well as a cluster analysis of the scientific publications, are important for defining promising areas of scientific research, which are being formed or actively developed in the scientific environment. This will make it possible to manage financial and organizational components of the development of specific research areas more effectively at the level of the state. It will also contribute to supporting the areas that are crucial in the strategy of scientific and technological development, to attracting international grants, to the creation consortiums for research, etc.

## 2. Literature review and problem statement

We shall consider identification of areas of research by scientists the process of establishment a correspondence between a specific scientist and scientific areas. The task on the identification of research areas by scientists has been actively addressed in the scientific environment. For example, in paper [1], this problem is solved by constructing methods for identifying areas of scientific activity based on a keyword analysis of publications. However, in paper [1], authors distinguish the publications that contain common keywords and phrases, but do not take into account other possible links between these publications: citation, similarity in the content of articles, etc. Article [2] proposes methods for employing the bibliometric attributes of scientific papers in order to cluster the authors of these papers by the areas of scientific research. Article [2] also considers a problem on determining new directions for scientific research by constructing a co-citation graph among the publications of scientists from different countries. Paper [3] described a technology for the categorization of objects by the adapted method of classification, which uses a special function of distance that makes it possible to take into account several attributes of the object. This technology enables parametric control over division according to the significance of attributes. This approach is important for the classification of various social phenomena. The method that is described in paper [3] can be also used for the problem on the distribution of researchers by area of research.

The task that is associated with the problem of identification of research areas is to identify groups of authors who work in a particular academic direction. A method for identifying groups of scientists based on an analysis of the similarity of texts of scientific papers was tackled in article [4]. However, study [4] failed to take into account that certain scientific direction can be shared by several groups of authors.

Paper [5] describes a method for determining the degree of similarity of scientific texts based on the method of locally-sensitive hashing for finding incomplete duplicates in the scientific articles. This task can be used to establish similarities between publications at the clustering stage of these publications. Article [6] presents a conceptual model of the automated system of finding incomplete duplicates, which is used for the implementation of methods outlined in paper [5]. This conceptual model can be used to implement a method of clustering of scientific publications, which is based on an analysis of similarities in the content of these publications.

To perform a cluster analysis of publications, it is convenient to employ a co-citation graph. Paper [7] considers methods for intelligent data mining that are represented by means of graphs, in particular, clustering of the co-citation graph. The task on the graph clustering has been studied rather sufficiently and there are many methods to solve this problem. One of the algorithms for solving a graph clustering problem is the algorithm Louvain, which is described in article [8]. This algorithm implements a method for maximizing graph modularity and can be used for rapid clustering of graphs with large dimensionality. This algorithm may be useful for clustering the co-citation graphs as these graphs are characterized by a considerable number of vortices and arcs. When conducting a graph clustering procedure, it is necessary to take into account the problem of insufficient stability of the clusters whose structure changes over time. Paper [9] proposed a method for determining cumulative cores, the use of which improves stability of clustering. Article [10] also examined the issues of stability of the graph clustering and proposed another method to resolve this problem, namely, finding stable alliances by sequential clustering of the graph. Study [11] suggested using the method of data clustering BIRCH (Balanced Interval Reducing and Clustering using Hierarchies) that makes it possible to minimize the number of requests to the database. Its application is appropriate when processing large volumes of data, in particular in the case when a database contains a considerable number of publications that have to be clustered.

Devising directions of scientific activity, within which scientists collaborate, based on the clustering of publications and identification of areas of research by scientists, is an important task for scientific and educational institutions. This task can be used to assess the activities of research institutions and to predict development prospects of organizations taking into account available resources. Paper [12] addresses the issue of constructing such models. In article [13], authors built a parametrical model that enables forecasting and assessment of the functioning of scientific and educational institutions based on the reallocation of existing resources. Paper [14] gives an overview of traditional methods for evaluating the results of scientific activity of researchers based on the analysis of citation of publications of these scientists. Paper [14] also proposes the scalar and vector methods for the evaluation of results of scientific activity. These methods can be applied for the assessment of results of scientific activity of researchers working in a certain scientific direction and the appropriate scientific field as a whole.

## 3. The aim and objectives of the study

The aim of present study is to conduct grouping of scientists in order to identify areas in which groups of scientists closely collaborate, using results of the scientific activity of these scientists: publication activity, citing of publications, etc. Identification of areas of research by scientists will make it possible to assess the contribution of each of them in the development of appropriate direction. Results of the identification would allow us to assess the direction as a separate object by analyzing the history of its development.

To accomplish the aim, the following tasks have been set:

– construction of the method for clustering the publications by scientists;

– construction of the method for identifying research results of scientists based on the results of clustering the publications of these scientists.

## 4. A method for the clustering of publications of scientists by scientific areas

Let $A=\{a_1, a_2,..., a_n\}$ be a certain set of scientists, $n$ is the number of scientists, $P=\{p_1, p_2,..., p_m\}$ is the set of publications that were published by the given scientists, $m$ is the number of publications. Let us also assume that there is a certain assigned metric space, which is a pair of $(P, g)$, where $P$ is the set of publications that were published by scientists from set $A$, g is the distance between the elements of set $P$, which is determined as a mapping from the set of Cartesian root of set $P$ onto the set of real numbers, that is:

$$g\colon P\times P\to \mathbf{R}, \tag{1}$$

where **R** is the set of real numbers.

In other words, the distance between publications is such an integral function $g(p_i,p_j)\geq 0,\ i=\overline{1,m},\ j=\overline{1,m},$ which is determined for arbitrary $p_i,p_j\in P$. In this case, for function $g(p_i, p_j)$, the axiom metrics are fulfilled: the axiom of identity (2), the axiom of symmetry (3), and the triangle inequality (4):

$$g(p_i,p_j)=0\Leftrightarrow i=j, \tag{2}$$

$$g(p_i,p_j)=g(p_j,p_i),\ \ \forall\, p_i,p_j\in P, \tag{3}$$

$$g(p_i,p_j)\leq g(p_i,p_e)+g(p_e,p_j),\ \ \forall\, p_i,p_j,p_e\in P, \tag{4}$$

Consider techniques to determine the distance between publications.

*A technique for determining the distance taking into account citations between publications.* Let the set $C\subset P\times P$ specifies a relation of citations between the publications by scientists. The relationship between publications and their citations can be represented in the form of a directed graph $(P, \mathrm{C})$, where publications from set $P$ are vertices, citations $C$ are the arcs of the graph. Fig. 1 shows a form of such a graph whose set P consists of 11 publications.

It is known that the route between vertices $p_i$ and $p_j$ in the graph denotes a sequence of such vertices

$$p_i=p_{i_0},p_{i_1},...,\ p_{i_u}=p_j,\ i_0<i_1<...<i_u,$$

$u\in\mathrm{N}$ each of which, except for the last, is connected to the next vertex with an edge or an arc. The length of the route is equal to the number of arcs in this route. Accordingly, the minimal route between the specified vertices is the route of the smallest length [7].

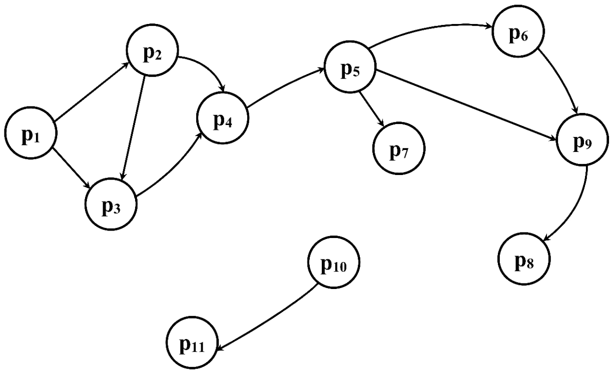

Fig. 1. Graph ($P$, $C$), which assigns a connection between publications $p_1$, $p_2$,..., $p_{11}$ and citations (graph arcs)

To set the distance between publications, it is possible to use the length of the minimal route between appropriate vertices of graph (P, C). If there is no route between the vertices, then the distance will be considered as an arbitrarily large.

*A technique for determining the distance between publications based on the degree of closeness by the content of abstracts of these publications.* To determine the degree of closeness of publications by the content, we shall compare not the entire text, but only the abstracts of these publications. This technique is based on the assumption that the abstracts of scientific publications that relate to the same scientific direction will contain the same concepts and keywords, that is, they will be quite similar in content. The method for determining closeness between fragments of text information, but when applied to the task on finding incomplete duplicates, was described in papers [5, 6].

In contrast to a full analysis of publications' text, comparison of abstracts for the purpose of determining the distance between publications has the following advantages:

1. Abstracts are much smaller in volume, which is why finding the distance between abstracts requires fewer computations.

2. Abstracts can be obtained from a bigger number of open sources than the full text of the publication.

3. An abstract contains information mainly in the text form. While content of the scientific publications may include tables, mathematical formulae, graphics. The presence of a considerable number of such objects complicates the task of comparing the publications.

4. A significant part of the publication contains summary of the author's material that is unique. If one finds closeness between the full text of two publications on the same subject, then the magnitude of distance between these publications is likely to be large.

5. An abstract reflects the basic content of the scientific publication.

Let abstract $S$ be assigned, which is a piece of text that consists of words. The word is a certain sequence of characters that belong to a finite alphabet $\overline{A}$. Denote words through $S_n^\beta,\ S_n^\beta\in S,\ n\in\mathrm{N}$ – a serial number of the word, β is the length of the word. Then an arbitrary word is assigned in the form:

$$S_n^{\beta}=\left\{t_1,t_2,\ldots,t_{\beta}\right\}, \qquad (5)$$

where $t_j \in \overline{A}$, $t_j \notin \overline{C}$, $j=\overline{1,\beta}$, $\overline{C}$ are all the non-character symbols.

We shall assign a list of stop-words and build the sequences of words from abstract S in the canonized form, that is,

$$W=\left\{S_1^{\beta_1},S_2^{\beta_2},\ldots,S_w^{\beta_w}\right\}, \qquad (6)$$

where $\beta_j$, $j=\overline{1,w}$ are the lengths of words, and w is the number of words.

Next, based on the fluid method window [5], we shall construct a set of sequences

$$F(W)=\{E_1, E_2,\ldots, E_{w-\omega+1}\}, \qquad (7)$$

which consist of a fixed number of words in the canonized form ω, $\omega<w$. And each subsequent sequence is built from the preceding with a shift by one word, which is chosen from sequence $W$.

We shall transform a set of sequences F(W) so that each sequence $E_i$ is represented by uniquely bit string $i=\overline{1,w-\omega+1}$. This can be done by using the method of locally-sensitive hashing. As a result, we shall obtain a set of bit strings

$$\Delta(W)=\{I(E_1), I(E_2),\ldots, I(E_{w-\omega+1})\}, \qquad (8)$$

where I(E$_i$) is the index element that specifies a bit string that uniquely represents sequence E$_i$, $i=\overline{1,w-\omega+1}$.

Each element of index I(E$_i$) will take the form:

$$I(E_i)=\{\delta_{i1}, \delta_{i2},\ldots, \delta_{ic}\}, \qquad (9)$$

where $\delta_{ix}\in\{0,1\}$, $i=\overline{1,w-\omega+1}$, $x=\overline{1,c}$, $c$ is the number of bits that forms the bit sequence $I(E_i)$.

Let $S^*=\{S_1, S_2,\ldots, S_m\}$ be a set of abstracts of scientific publications $p_1$, $p_2$,..., $p_m$. For each element of the set $S^*$ for $q=\overline{1,m}$ we shall construct a set of index elements by formula (8). As a result, we shall obtain:

$$\Delta\left(W^q\right)=\left\{I\left(E_1^q\right),I\left(E_2^q\right),\ldots,I\left(E_{w-\omega+1}^q\right)\right\}, \qquad (10)$$

where $\Delta\left(W^q\right)$ is the set of index elements for each abstract from set $S^*$.

Using a method of locally-sensitive hashing, in accordance with the method described in paper [5], we shall determine index elements:

$$I\left(E_i^q\right)=\left\{\delta_{i1}^q,\delta_{i2}^q,\ldots,\delta_{ic}^q\right\}, \qquad (11)$$

where $\delta_{ix}^q\in\{0,1\}$, $i=\overline{1,w-\omega+1}$, $x=\overline{1,c}$, $q=\overline{1,m}$.

The distance between abstracts in this case can be calculated based on the Hamming distance between all elements of each index of these abstracts. Thus, the distance between two given abstracts $S_\sigma$ and $S_\tau$ is calculated from formula:

$$H\left(S^{\sigma},S^{\tau}\right)=\frac{1}{c(w-\omega+1)}\sum_{i=1}^{w-\omega+1}\sum_{x=1}^{c}\left|\delta_{ix}^{\sigma}-\delta_{ix}^{\tau}\right|, \qquad (12)$$

where $H\left(S^{\sigma},S^{\tau}\right)$ is the Hamming distance between abstracts $S_\sigma$ and $S_\tau$, $\delta_{ix}^{\sigma}$ and $\delta_{ix}^{\tau}$ are the bits of index elements of the corresponding sets $\Delta\left(W^{\sigma}\right)$ and $\Delta\left(W^{\tau}\right)$, $\sigma\neq\tau$,

$$\sigma\in\{1,2,\ldots,m\},\ \ \tau\in\{1,2,\ldots,m\}.$$

The distance between publications will be considered equal to the distance between the abstracts of these articles, that is,

$$g(p_\sigma,p_\tau)=H\left(S^{\sigma},S^{\tau}\right), \qquad (13)$$

where $S_\sigma$ and $S_\tau$ are the abstracts of publications that are compared, $\sigma\neq\tau$.

It should be noted that for the axioms of a metric to hold, it is required that the additional constraint is satisfied: we shall assume that there are no two different abstracts $S_\sigma$ and $S_\tau$ with a zero Hemming distance, that is, if $\sigma\neq\tau$, then $H\left(S^{\sigma},S^{\tau}\right)\neq 0$. In fact, various publications with identical abstracts cannot exist. Coincidence of abstracts for various publications can be considered a case of plagiarism.

After we determined a technique to calculate distances between publications, it is possible to pass over to the procedure of the clustering of these publications.

*Statement of the problem on clustering the publications.* Let a metric space ($P$, $g$) be assigned. It is required to split the set of publications $P$ into certain number of subsets that do not intersect. Such subsets of the set $P$ are called clusters. Denote the set of clusters through $Y=\{y_1, y_2,\ldots, y_z\}$, where $z$ is the number of clusters into which set $P$ is split. In order to assign a set of clusters Y, the following conditions must be met:

1. Each publication necessarily belongs to one of the clusters, that is,

$$\bigcup_{i=1}^{z} y_i = P.$$

2. Each publication belongs to a single cluster, that is, $y_i\cap y_j=\varnothing$, $\forall\, i\neq j$.

3. Each cluster covers sufficiently close publications (in the sense of distance $g$).

It should be noted that in the case of calculation of distances between publications applying the above-described techniques, there is a possibility for the occurrence of the so-called isolated publications. A publication will be considered isolated if the distance between it and any other publication from the set P is infinite.

According to the definition of a cluster, no other publication may belong to the cluster that the isolated publication belongs to. This leads to the emergence of clusters consisting of only a single publication. The presence of such clusters complicates further processing of the data and does not contain any information concerning the direction of scientific research of the isolated publications. That is why, in order to perform the clustering procedure, it is proposed to exclude the isolated publications from consideration.

*Methods for clustering the publications.* We shall consider methods that can be employed for conducting the clustering procedure of scientific publications with the assigned techniques for determining a distance between these publications.

It is possible to select a separate class of the clustering methods, which are used in the case of data representation in the form of a graph. The main feature of graph clustering is the discrete metric space. The distance can be determined only between vertices of the graph and it is impossible to determine the distance from any arbitrary point that does not belong to the graph. Because of this, many classical clustering methods such as c-means, PAM, Hierarchical methods, SOM [15], and others, are not applicable for the clustering of graphs that represent links between scientific publications and citations. In order to cluster such graphs, it is proposed to use specialized clustering techniques, including the Louvain method. The Louvain method is based on maximizing the modularity of the graph. Modularity is a numerical estimation of the quality of graph's splitting into subgraphs. Modularity is defined as the sum of differences between the parts of arcs Doug in the corresponding subgraph and the squared parts of arcs whose one end belongs to the appropriate subgraph. In other words, modularity of the clustering of publications graph can be determined as:

$$Q = \sum_{v=1}^{z} (\beta_v - \alpha_v^2), \qquad (14)$$

where $\alpha_v$ is the share of citations, where the publication that contains this citation, or the publication which was cited, belongs to cluster $y_v$, that is,

$$\alpha_v = \frac{\left\| \left\{ p_i \in P \middle| (p_i, p_j) \in C, p_i \in y_v, p_j \in P \right\} \cup \left\{ p_i \in P \middle| (p_j, p_i) \in C, p_i \in y_v, p_j \in P \right\} \right\|}{\mathrm{card}(C)},$$

while $\beta_v$ is the share of citations, where the publication that contains this citation and the publication that was cited belongs to cluster $y_v$,, that is,

$$\beta_v = \frac{\left\| \left\{ p_i \in P \middle| (p_i, p_j) \in C, p_i \in y_v, p_j \in y_v \right\} \right\|}{\mathrm{card}(C)},$$

where card$(C)$ is the number of arcs of graph $(P, C)$.

The first step of the Louvain method is the initial splitting of the graph so that each vertex of the graph forms a separate cluster. This splitting is matched by the minimum value of modularity. Next, we perform iterative procedure for merging the clusters. Each such merger is matched by the maximum increase in the modularity of the graph. Merging the clusters is carried out as long as it is possible to increase the modularity of the graph. To find a solution to this optimization problem, one of the simulation methods is applied, such as the Monte Carlo method.

A special feature of the Louvain method is that it does not use neural networks, which is why it requires training samples. However, the method has proved to be effective for the clustering of graphs with large dimensionality [8].

One of the problems that can arise in the course of implementation of the procedure of graph clustering is the stability of splitting the graph into clusters. This problem occurs when there are several techniques for splitting, which are matched by close values of modularity. To solve this problem, it is proposed to apply a known method for finding stable assemblies of the graph's vertices $(P, C)$, which is described in paper [10]. Fixing such assemblies makes it possible to build robust clusters of publications, that is, it solves the problem on the stability of clustering.

We shall assume that the initial graph $(P, C)$ was clustered, for example, based on the Louvain method, and we received initial partition of the set of publications P into z clusters. We selected $z$=3 clusters $y_1, y_2, y_3$ in the graph shown in Fig. 1. These clusters are marked with colors in Fig. 2.

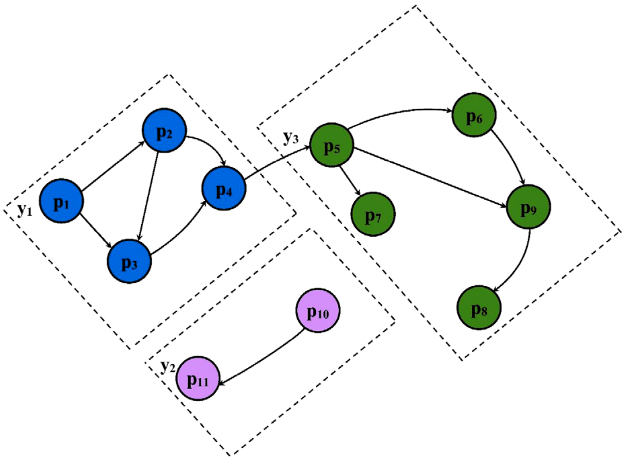

Fig. 2. Results of the clustering of graph $(P, C)$ into 3 clusters $y_1, y_2, y_3$

*An analysis of results of the clustering of scientific publications.* As a result of the procedure of clustering of scientific publications, we obtained a set of clusters *Y*. The power of set Y can be quite large, making it difficult to run further analysis. One technique to solve this problem is the enlargement of constructed clusters by merging the clusters that are close together, containing a small number of elements. For this purpose, it is necessary to determine the centre of gravity of each built cluster.

The center of gravity of cluster

$$y_k = \left\{ p_1^k, p_2^k, \ldots, p_{\mu_k}^k \right\}, \quad k = \overline{1, z},$$

is such an object from cluster $\Omega^k$ that the total distance to other objects of this cluster is minimal:

$$\Omega^k = \arg\min \left( \sum_{i=1}^{\mu_k} g(p_i^k, p_j^k), j = \overline{1, \mu_k} \right), \qquad (15)$$

where $\mu_k$=card$\underline{(y_k}$) is the number of objects that belong to cluster $y_k$, $k = \overline{1}, z$.

Fig. 3 shows the result of merging the clusters, built by the procedure of clustering of graph $(P, C)$.

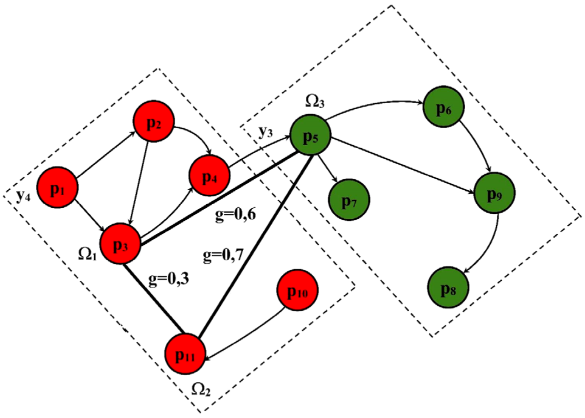

Fig. 3. Results of execution of the algorithm for merging the clusters built for graph $(P, C)$

The algorithm to implement such a merger of clusters consists of the following steps:

1. Set the counter $b$=0.

2. Find centers of gravity $\Omega^k$ of each cluster

$$y_k = \left\{ p_1^k, p_2^k, \ldots, p_{\mu_k}^k \right\}.$$

3. Calculate distances between the centers of gravity of each cluster. The basis can be formula (13), by comparing the abstracts of publications that match centers of gravity.

4. If there are such clusters $y_k$ and $y_l$ that the distance between their centers of gravity does not exceed the threshold value δ, that is, the condition $g\left(\Omega_k, \Omega_l\right) \leq \delta$, is satisfied, then:

4. 1. Increase the counter by unity: $b$=$b$+1.

4. 2. Form a new cluster $y_{z+b} = y_k \cup y_l$.

4. 3. Clusters $y_k$ and $y_l$ are excluded from further consideration.

4. 4. Find the center of gravity of the cluster $\Omega_{z+b}$.

4. 5. By using formula (13), we find distances from the center of gravity of cluster $y_{z+b}$ to the centers of gravity of other clusters $y_1$, $y_2$,..., $y_{z+b-1}$.

4. 6. Return to point 4.

5. If the distance between the centers of gravity of clusters exceeds the threshold value for all clusters, that is, there are no such $y_k$ and $y_l$ for which condition $g\left(\Omega_k, \Omega_l\right) \leq \delta$, is satisfied, then the execution of the algorithm ends.

It should be noted that there may be scientific publications, which obtained as a result of research that is actually performed in a joint scientific direction, but in terms of the proposed distances, these publications will be far apart. There are many examples when several groups of scientists did obtain the same results independently of each other. In this case, the publications of respective outcomes will not include cross references and will be distant from each other if the distances are calculated taking citations into consideration. That is why, to solve this problem, it is proposed to calculate distances by both techniques. If, when clustering publications, the distance is determined based on citations, then when the clusters are merged, the distance should be found based on the closeness of abstracts' content and vice versa.

## 4. The task of the identification of scientists' research areas

*Establishing an alignment of the cluster, which consists of publications by scientists, to the field of research.*

After the clustering procedure of graph ($P$, $C$) and merging the clusters that are close to each other, it is required to establish alignment of the specific cluster to the verbal name of the research area represented by this cluster. That is, if $Y$={$y_1$, $y_2$,..., $y_z$} is the constructed set of clusters after applying one of the clustering algorithms of graph ($P$, $C$), and

$$\overline{Y} = \left\{ y_{k_1}, y_{k_2}, \ldots, y_{k_\psi} \right\}$$

is the ultimate set of clusters, which is built as a result of the execution of algorithm for merging close clusters,

$$k_j \in \left\{ 1, 2, \ldots, z, z+1, \ldots, z+\nu \right\}$$

are the indexes of elements of the ultimate set of clusters, $h = \overline{1, \psi}$, $\nu$ is the number of mergers of clusters during execution of the merging algorithm, ψ is the number of elements of the resulting set of clusters.

Each cluster $y_{k_1}, y_{k_2}, \ldots, y_{k_\psi}$ will be assigned with a certain direction of scientific research. That is, we shall consider mapping $\Phi: \overline{Y} \to V$, where $V$ is the set of verbal names of research fields. For example, the elements of set $V$ may include: "Mathematical Physics", "Theory of Optimization", "Computer Science", etc. To establish correspondence Φ, one can use an expert approach. In this case, experts will make a decision about establishing correspondence between each cluster and an appropriate field, based on the list of publications in the cluster and some additional information, such as keywords, the most widely used concepts, etc.

*Identification of the scientists' research areas.*

Let $A$={$a_1$, $a_2$,..., $a_n$} be a certain set of scientists, $n$ is the number of scientists, and $P$={$p_1$, $p_2$,..., $p_m$} is the set of publications that were published by the given scientists, m is the number of publications. Denote through $V$={$\eta_1$, $\eta_2$, ..., $\eta_\psi$} a set of areas of scientific research, ψ is the number of fields of research. As already noted, the identification of scientists' research areas is the process of establishing a correspondence between a particular scientist and the scientific areas in which this scientist is engaged and which are addressed in his/her publications within the framework of the given areas. That is, it is required to find representation Λ:*A*®*V*. To identify the areas of scientists' research, one technique is to use information on the publication activity of scientists, taking into consideration the built set of clusters of research areas to which these publications belong. It is clear that scientific publications in the vast majority are published with co-authors. In graph ($P$, $C$), which is shown in Fig. 2, 3 after clustering, graphic identification of the authors' research areas will take the form shown in Fig. 4.

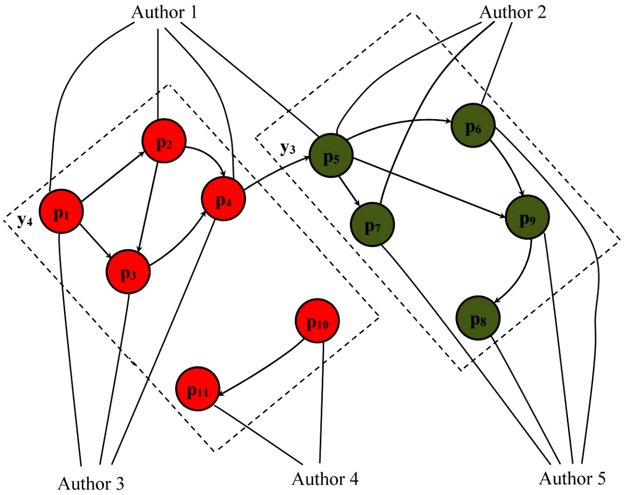

Fig. 4. Results of the identification of scientists' areas of research using graph (*P, C*) as an example

A set of research areas of scientist $a_i$ is determined as:

$$\Theta\left(a_i\right) = \left\{ \Phi\left(y_{k_h}\right) = \eta_h, h = \overline{1, \psi} \middle| y_{k_h} \cap P\left(a_i\right) \neq \text{Ш} \right\}, \qquad (16)$$

where $\Theta(a_i)$ is a set of research areas of scientist $a_i$,

$$i = \overline{1, n}, \quad \Phi: \overline{Y} \to V.$$

The totality of sets $\Theta(a_i)$ for all scientists is the solution to the problem of the identification of research areas. The obtained results make it possible to solve the related problem on finding the scientists that undertake research in the specified area.

## 5. Discussion of results of research into the clustering of scientific publications and the identification of scientists' areas of research

In the course of the present study we analyzed known methods for clustering the scientific publications and identifying the areas of research. The analysis revealed that many of them, including the methods described in papers [1, 2], are limited by the significant impact of the keywords selected in publications on the result of clustering. These methods can be applied locally to identify sub-areas in a certain area of research, but they are difficult to employ for a full-scale clustering of all publications, information on which is derived from scientometric databases.

In order to eliminate the specified shortcomings, and to identify scientists' areas of research, we devised an improved method for the clustering of scientific publications by appropriate areas. The task on clustering the scientists' publications is sometimes based on finding a distance between full texts of the publications. In particular, paper [4] describes the use of a full-text analysis of publications to identify groups of authors who carry out research in the field of physical-mathematical sciences.

In the authors' opinion, in order to find a distance between publications, it will suffice to compare their abstracts. Comparison of abstracts for finding the distance between publications is explained by the fact that in contrast to the analysis of abstracts, a full-text analysis of papers requires a large volume of computations and full access to the texts of scientific research. The average volume of scientific publications exceeds 3,000 thousand words while the average size of an abstract is 150 words. The difficulty of finding a distance using the locally-sensitive hashing is proportional to the number of words. Accordingly, finding the distance between abstracts is carried out, on average, by 30 times faster than between publications.

For the experimental analysis of the devised method, we compiled a database containing information about 215,082 publications by 58,834 Ukrainian authors. As a result of the application of the citation graph clustering method, it was possible to identify about 20 thousand clusters or scientific areas. When carrying out a procedure of merging the clusters it was discovered that a decrease in the number of clusters results in an increase in the instability of citation graph clustering. The reason for instability is insufficient connectedness of the graph. To solve this problem, it is necessary to complement input data with publications of foreign authors. The application of the method for merging the clusters, based on finding a distance between abstracts, reduces the number of clusters or areas of scientific research to about 500. This greatly simplifies further analysis and evaluation of these areas.

The use of each method separately leads to the occurrence of isolated clusters of publications that may belong to the same area of scientific research. In particular, there are groups of scientists who examine one topic in parallel, but there are few citations related to their publications, or the citations are absent at all. Similarly, given different style of writing by authors, there are abstracts for publications, which emphasize different aspects of the same problem, which is why they are quite far in terms of content. Applying both methods for estimating a distance between the abstracts makes it possible to identify publications belonging to a common area of research more accurately.

The main difficulties that arise in the identification of scientists' areas of research are:

1. Completeness of the information on citing the publications is a necessary condition for the implementation of the method. That is why there is a need to provide an access to scientometric databases that contain this information. Given the trends in the implementation of national scientometric databases in Poland, Russia, other countries, there is a problem related to adding the publications of foreign scientists for consideration. In the absence of such publications, citation graph (P, C) is incomplete, which is why there appear unlinked clusters of publications that affects the results of clustering of scientific publications.

2. The necessity of finding a locally-sensitive hash for the abstract of each examined publication and finding a matrix of distances between them. Each of these activities has considerable computational complexity. To reduce the load on the system, it is proposed to find a hash of the abstract when adding the publication to the system in order to save it as one of the characteristics of this publication.

3. Determining a threshold value of distance δ, at which clusters should be merged. The problem is that different fields are characterized by a different theoretical mean distance between scientific publications.

4. Building a match between clusters and areas of scientific research requires engagement of experts. During operation of the system, a database of the basic concepts of the field is created, which is based on keywords from the publications and terms that are most often found in the abstracts of these publications.

Considering the specified difficulties, in order to cluster scientific publications, it is more appropriate to use distances between publications determined on the basis of citation. Finding such distances requires a less amount of computations than finding distances between publications based on the closeness of publications by content. It should also be noted that the co-citation graph (P, C) is rather sparse: each vertex typically generates around 10–20 arcs. A graph built based on the distance between abstracts by content is almost complete, that is, the number of edges that emanate from each vertex is slightly less than the number of vertices in the graph. Clustering of the sparse graph is a simpler task than clustering of the full graph.

When merging the clusters, a distance is determined only between the centers of gravity of the clusters whose number is significantly less than that of publications in general. That is why at this stage it is more appropriate to apply a method of calculating the distances between abstracts of scientific publications. Under such a sequence of actions, the clustering of scientific publications requires a smaller amount of computations.

In the future, it is planned to devise a method for calculating a short-term forecast of gain in a change of integrated assessments of a scientific area. It is assumed that the method to be designed could be applied to identify promising areas of research.

## 8. Conclusions

1. We constructed a method for the clustering of scientific publications by research areas. The method uses representation of publication activity and citation of authors

in the form of a directed graph. The method employs several proposed techniques to calculate a distance between publications: based on the degree of closeness by the content of abstracts for these publications, and by considering citation links between publications. As a result, we built a set of clusters, each of which contains a certain number of scientific publications that are close to each other. Due to the specificity of input data, in the process of clustering there may emerge the isolated publications and the clusters that are close enough, which contain a small number of publications. That is why we proposed an algorithm for merging the clusters that are close, and the exclusion of the isolated publications from consideration.

2. To identify scientists' areas of research, it is proposed to initially establish a correspondence between the clusters and the appropriate verbal representations of research areas by using expert methods. After that, it becomes possible for each scientist to form a set of areas for scientific research, taking into account the mapping of a set of publications by scientists onto a set of scientific areas.

## References

1. Bhattacharya, S. Mapping a research area at the micro level using co-word analysis [Text] / S. Bhattacharya, P. K. Basu // Scientometrics. – 1998. – Vol. 43, Issue 3. – P. 359–372. doi: 10.1007/bf02457404
2. Glänzel, W. Bibliometric methods for detecting and analysing emerging research topics [Text] / W. Glänzel // El Profesional de la Informacion. – 2012. – Vol. 21, Issue 2. – P. 194–201. doi: 10.3145/epi.2012.mar.11
3. Mulesa, O. Information technology for determining structure of social group based on fuzzy c-means [Text] / O. Mulesa, F. Geche, A. Batyuk // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). – 2015. doi: 10.1109/stc-csit.2015.7325431
4. Shvets, A. Detection of current research directions based on full-text clustering [Text] / A. Shvets, D. Devyatkin, I. Sochenkov, I. Tikhomirov, K. Popov, K. Yarygin // 2015 Science and Information Conference (SAI). – 2015. doi: 10.1109/sai.2015.7237186
5. Lizunov, P. Detection of near dublicates in tables based on the locality-sensitive hashing method and the nearest neighbor method [Text] / P. Lizunov, A. Biloshchytskyi, A. Kuchansky, S. Biloshchytska, L. Chala // Eastern-European Journal of Enterprise Technologies. – 2016. – Vol. 6, Issue 4 (84). – P. 4–10. doi: 10.15587/1729-4061.2016.86243
6. Biloshchytskyi, A. Conceptual model of automatic system of near duplicates detection in electronic documents [Text] / A. Biloshchytskyi, A. Kuchansky, S. Biloshchytska, A. Dubnytska // 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM). – 2017. doi: 10.1109/cadsm.2017.7916155
7. Samatova, N. Practical Graph Mining with R [Text] / N. Samatova, W. Hendrix, J. Jenkins, K. Padmanabhan, A. Chakraborty. – Chapman and Hall/CRC, 2013. – 495 p.
8. Blondel, V. D. Fast unfolding of communities in large networks [Text] / V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre // Journal of Statistical Mechanics: Theory and Experiment. – 2008. – Vol. 2008, Issue 10. – P. P10008. doi: 10.1088/1742-5468/2008/10/p10008
9. Seifi, M. Community codes in evolving networks [Text] / M. Seifi, J.-L. Guillaume // Proceedings of the 21st international conference companion on World Wide Web – WWW '12 Companion. – 2012. – P. 1173–1180. doi: 10.1145/2187980.2188258
10. Ovelgönne, M. An ensemble learning strategy for graph clustering [Text] / M. Ovelgönne, A. Geyer-Schulz // Contemporary Mathematics. – 2013. – Vol. 588. – P. 187–205. doi: 10.1090/conm/588/11701
11. Zhang, T. BIRCH: an efficient data clustering method for very large databases [Text] / T. Zhang, R. Ramakrishnan, M. Livny // Proceedings of the 1996 ACM SIGMOD international conference on Management of data. – 1996. – Vol. 25, Issue 2. – P. 103–114. doi: 10.1145/233269.233324
12. Otradskaya, T. Development process models for evaluation of performance of the educational establishments [Text] / T. Otradskaya, V. Gogunsky // Eastern-European Journal of Enterprise Technologies. – 2016. – Vol. 3, Issue 3 (81). – P. 12–22. doi: 10.15587/1729-4061.2016.66562
13. Otradskaya, T. Development of parametric model of prediction and evaluation of the quality level of educational institutions [Text] / T. Otradskaya, V. Gogunskii, S. Antoshchuk, O. Kolesnikov // Eastern-European Journal of Enterprise Technologies. – 2016. – Vol. 5, Issue 3. – P. 12–21. doi: 10.15587/1729-4061.2016.80790
14. Biloshchytskyi, A. Evaluation methods of the results of scientific research activity of scientists based on the analysis of publication citations [Text] / A. Biloshchytskyi, A. Kuchansky, Y. Andrashko, S. Biloshchytska, O. Kuzka, O. Terentyev // Eastern-European Journal of Enterprise Technologies. – 2017. – Vol. 3, Issue 2 (87). – P. 4–10. doi: 10.15587/1729-4061.2017.103651
15. Jain, A. K. Data clustering: a review [Text] / A. K. Jain, M. N. Murty, P. J. Flynn // ACM Computing Surveys. – 1999. – Vol. 31, Issue 3. – P. 264–323. doi: 10.1145/331499.331504