Для задачі формування проектних груп, зокрема науково-дослідницьких проектних груп, пропонується комплексний метод, який складається з двохетапного методу кластеризації графу цитування публікацій науковців та методу нечіткого логічного виводу для узгодження думок експертів щодо вибору потенційних партнерів і включення їх до проектної групи.

Суть двохетапного методу кластеризації публікацій науковців полягає у кластеризації графу цитування та об'єднання кластерів на основі близькості анотацій публікацій. Відстань між публікаціями розраховується на основі визначеної метрики та підходів n-грам аналізу. Описаний метод дозволяє ідентифікувати напрями досліджень науковців, що є необхідною складовою раціонального вибору партнера для побудови проектної групи і є вхідною інформацією для експертів, які цю групу формують. Наступним етапом є застосування методу нечіткого логічного виводу, який будується для узгодження думок експертів щодо створення проектних груп. Даний метод складається із трьох етапів. На першому етапі фазифікація здійснюється через введення функції належності науковця до напрямку наукових досліджень. Другий етап нечіткого логічного виводу полягає формуванні експертами вимог до кандидатів на місце в проектній групі. На заключному етапі відбувається дефазифікація за допомогою методу центра ваги. Для верифікації нечіткого методу ідентифікації дослідницьких проектних груп було визначено організації-виконавці для фундаментального наукового дослідження.

Описані методи можуть бути використані для задачі формування науково-дослідницьких груп та виявлення подібностей між фрагментами текстової інформації на основі n-грам аналізу, що має застосування у задачі ідентифікації неповних дублікатів між фрагментами текстової інформації

Ключові слова: кластеризація, n-грам аналіз, напрям наукових досліджень, граф цитування, проектна група

UDC 005.8
DOI: 10.15587/1729-4061.2019.175139

IMPROVEMENT OF THE METHOD FOR SCIENTIFIC PUBLICATIONS CLUSTERING BASED ON N-GRAM ANALYSIS AND FUZZY METHOD FOR SELECTING RESEARCH PARTNERS

P. Lizunov
Doctor of Technical Sciences, Professor, Head of Department
Department of Computer Science*
E-mail: lizunov@knuba.edu.ua
A. Biloshchytskyi
Doctor of Technical Sciences, Professor, Head of Department**
E-mail: bao1978@gmail.com
A. Kuchansky
PhD, Associate Professor**
E-mail: kuczanski@gmail.com
Yu. Andrashko
PhD, Associate Professor
Department of System Analysis and Optimization Theory
Uzhhorod National University
Narodna sq., 3, Uzhhorod, Ukraine, 88000
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
*Kyiv National University of Construction and Architecture
Povitroflotskyi ave., 31, Kyiv, Ukraine, 03037
**Department of Information Systems and Technologies
Taras Shevchenko National University of Kyiv
Volodymyrska str., 60, Kyiv, Ukraine, 01033

Received date 15.05.2019
Accepted date 24.07.2019
Published date 29.08.2019

1. Introduction

Under modern conditions, the problem of increasing the effectiveness of scientific research and establishing effective cooperation within scientific communities is becoming increasingly acute. Creation of teams that are able to successfully implement projects is especially important for project-oriented organizations: scientific research institutions, consulting companies, design bureaus, etc. The typical approach to the formation of the scientific research project team is to choose partners from a range of scientists, who have relevant qualification and experience of realization of

similar projects. That is, the selection is often based on analysis of previous experience and scientific results of potential project partners.

Scientific communities are usually formed under conditions of competition. In the situation of globalization of scientific research, not only the space of one or more countries, but rather of the whole world is taken into account. These scientific communities may consist of universities, scientific research institutes, private companies, foundations and associations that form the applications for joint grants. Since only a part of the grants, which were applied for, pass and receive funding, it is possible to distinguish three forms of relationships for every two subjects of scientific communities (universities, institutes and companies): competition, neutral relationship and partnership. Such relationships may develop not only between entities of scientific communities, but also between their separate structural units (departments, faculties, scientific research departments, project teams, etc.).

The problem of choosing partners in the framework of a grant or a strategic partner for the implementation of a series of scientific research is relevant, especially under conditions of globalization and intensive development of mobility of scientific communities. One of the approaches that can be used to solve this problem is to conduct clustering of scientific publications of researchers, who are potential partners in a project. In addition, to support the rational choice, it is necessary to conduct analysis of citations between these publications and establish proximity between publications based on the *n*-gram analysis of abstracts. Scientifically grounded methods for choosing a partner are the key to creating and effective development of a scientific research project in future. Thus, the problem of choosing partners in the formation of research project groups is a relevant task of research.

2. Literature review and problem statement

Analysis of the relations between documents provides considerable information about the structure of the space of documents, which can be used for a wide range of practical tasks. Methods for clustering scientific publications can be used to identify the areas of scientific research. In paper [1], there is a comparison of the methods for clustering of citation graph: the spectral method [2], the method for modularity maximization [3, 4], matrix factorization [5], and the method for drawing up the Map of random walks [6].

The result of the study, which is described in [1], shows that the quality of the construction of clear clusters of documents based on analysis of relations graph strongly depends on graph density, that is, qualitative clustering is possible only in the presence of a sufficient number of related documents in the collection. Among 11 algorithms considered in this research, the Louvain method expectedly became the best by the criterion of modularity [4]. The Louvain method allows constructing the least number of clusters. By the Flake function criterion, which shows the ratio of the number of vertices that have higher external power to the number of vertices that have higher inner power in a cluster, the Louvain method took second place, being slightly inferior to the BPA. And finally, by the time criterion, only three methods: the Louvain, Walktrap and BPA methods can be used for clustering of very large graphs. The authors of the research made a comparison of results of the explored clustering methods with the expert method, and came to the conclusion that all the explored methods are significantly inferior to the expert method. Therefore, the problem of the research is to verify the effectiveness of application of combined methods based on clustering of citation graph taking into account additional information (keywords, the title of the article, etc.) to the problem of identifying the areas of scientific research.

Paper [7] presented the information-theoretical approach to finding the distribution of weight of feature space. The use of feature space to increase the original similarity of documents allows achieving their better clustering. Thus, the construction of the complex methods that, along with the citation graph, takes into account other properties of publications, makes it possible to improve the clustering quality. However, the specific features of practical realization of the approach were not revealed in the work to full extent.

Paper [8] contains the classification of the methods for finding similarities of texts. By the subject of comparison, all methods for determining the similarity of the text are divided into String-Based, Corpus-Based and Knowledge-Based methods. Among the String-Based methods, it is possible to separate the *n*-gram method of analysis. Results of study [8] indicate the possibility of using the method of *n*-gram analysis for the problem of identifying the similarities of a publication. The identified similarity can be additional information when carrying out clustering of scientific publications. In study [9], there was an analysis of using *n*-gram at $n \in \{1, 2, 3, 4, 5\}$ for finding of similarity of the text based on the use of a match of search queries of Google Web 1T data set [10]. It was shown that the application of the three-gram ($n=3$) has an advantage over the use of bi-gram ($n=2$), 4-gram ($n=4$) and 5-gram ($n=5$). The application of unigram ($n=1$) is a degenerated case and is a simple check of words entering, so is the least effective.

The problem of the formation of scientific project groups under conditions of globalization and mobility of scientific communities requires new application of new approaches to its solution. In this context, we can highlight the problem of calculating the rating of competitors, evaluation of the activity of other companies and institutions, which can potentially become partners. To find the ratings of entities of scientific communities and separate scientists, it is possible to use the principle of construction of metrics. In this case, it is important to take into account the fluidity and dynamism of indicators of activity of entities of scientific communities and separate scientists, as well as the specifics of partnership activity and specificity of already formed competitors. A promising direction in the development of technologies of evaluation of activity results of entities of scientific communities is construction of integrated estimates, which are described in papers [11, 12]. The unresolved issue is the use of constructed estimates for the problem of choosing research partners.

Articles [13–15] presented the ways of project management integration and decision-making support by using a matrix model based on the key portfolio events. The unresolved issue is the use of the matrix model for the problem of choosing research partners. In the paper [16], partnership relationships and principles of their management were studied. Study [17] described the multi-stage mathematical model of choosing partners. In paper [18], it was proposed to use the method of analytical hierarchy to select the best partners from a small set of potential partners. Article [19] described the model of choosing the partner based on the genetic algorithm. Paper [20] proposed the approach to the construction of the information technology based on

the project-vector methodology for the scientific research management. The development of the information [21] and technical [22] component of the technology shows the prospects of its application. In the explored works, the models of choosing partners are mostly based on the idea that potential partners can perform a wide range of functions, but in actual projects, especially scientific research projects, it is not so. Each partner can perform narrow tasks, which are a part of the road map of this project development. That is why the models and the methods, which are described in these works, require revision.

While selecting research partners, the considered methods take into account only their results obtained in the past. To determine a rational partner, it is important to assess the prospects of his development as a subject, at this, appropriate mathematical methods of forecasting and expert evaluation should be applied. In paper [23], the theoretical approach to applying expert methods when predicting the development of subjects was used. The proposed models can be used in the construction of time series of characteristics, which are used for the evaluation of the prospects of a partner. Article [24] shows the method for prediction of time series of selective match with the model, which was developed in [25] and integrated into information technology in [26]. These studies show the application of trend models for the prediction of economic processes, while paper [27] shows the possibility of application of trend models for forecasting the prospects of scientific research. It was shown that the trend methods best predict the prospects of scientific research areas The prospects of separate scientists and scientific research institutions are predicted with less accuracy, which is why the choice among the forecasting models and methods of the ones, which would more accurately predict the potential of scientists and scientific research institutions [28], remains an unresolved problem.

Paper [29] deals with the problem of selecting the project executor as a problem of multicriteria decision making. The proposed method is based on the process of hierarchy analysis. To solve the obtained optimization problem, the method of lexicographic target programming with the use of Boolean variables is proposed. Given the *NP* complexity of the obtained problem, this approach can be used at a small number of potential project executors.

In study [30], the problem of selecting executors for the implementation of individual tasks of a new project was considered. Two approaches to problem solving were proposed. The first approach allows taking into account the assessment of the competence of potential executors and the data on already realized projects. These data and requirements for the tasks of new projects are presented using the cognitive map with corresponding vertices. The major difficulty of this method is the need for constructing a graph of relations between executors, projects and tasks because of its large dimensionality. A rather challenging task is determining numerical parameters of a cognitive map: impact weights, importance of the tasks of a new project, etc.

The second approach involves using the Hopfield neural network. Neural networks allow the implicit use of already existing experience of project activities during the implementation of new projects. The main problem of the application of neural networks is the need for long training and selection of training samples.

Study [31] focuses on the research into increasing the effectiveness of the formation of a project team with regard to interpersonal relationships of employees. The criterion for determining the contribution of employees to the group interaction was proposed. Using this criterion, the algorithm of solving the problem of selection of candidates for the project team by the criterion of total contribution of employees to the group interaction was developed. Paper [32] shows that team with different level of integration of executors had the same levels of effectiveness of teamwork. Thus, the data obtained as a result of the study indicate that the role and importance of integration in project groups is insignificant compared to other approaches to the productivity improvement, in particular, the choice of executors.

Thus, after examining these studies, we can come to the conclusion that the problem of selection of scientific partners can be solved more efficiently, taking into account the specifics of scientific research projects, which are often performed by researchers of narrow specialization. To enhance the efficiency of determining the specialization a researcher based on his previous experience, there is a need for improvement of the methods for identification of the research areas, in which a scientist performed before.

3. The aim and objectives of the study

The aim of this study is to improve the method of clustering publications of scientists based on criteria of abstract proximity and relations between citations. The results of the method will give an opportunity to identify more clearly the research project groups for solving the problem of selecting research partners.

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

– to construct a method for determining the proximity of abstracts based on the *n*-gram analysis for the scientists publications clusterization problem;

– to select the methods of fuzzification of potential partners matching the research problems of a project according to the results of clustering and defuzzification of evaluations of partners based of the requirements stated by experts;

– to conduct verification of obtained results.

4. Construction of the method for determining the proximity of abstracts based on the *n*-gram analysis for the scientists publications clusterization problem

In paper [33], the identification of research areas of scientists is determined as a process of finding a match between a specific scientist and scientific areas, in which this scientist works and publishes his scientific papers in the framework of these areas. The method for identification of research areas of scientists that consists of four states was proposed:

1. Clustering of publication citation graph.

2. Consolidation of clusters.

3. Construction of the match between clusters of publications.

4. Identification of research areas of scientists.

At stage 1, the set of publications $P=\{p_1, p_2, ..., p_m\}$ as vertices of the citation graph, which are connected with each other by arcs – citations, is considered. As a result of the procedure of clustering scientific publications, the set of clusters Y is obtained. Since the power of set Y can be quite large, the need for the aggregation of the constructed clusters

by merging the clusters that are close to one another with a small number of elements is shown. The paper proposed the method for determining the distance between publications based on determining content proximity of abstracts by content. It was proposed to find the distances between abstracts using the method of locally-sensitive hashing. As practice shows, the use of this method gives satisfactory results when comparing the abstracts written by scientists from one region. Cultural, regional and language traditions of the authors have significant influence on the writing style, the use of certain words and phrases. When comparing the abstracts of scientists from different countries, there arises a situation when the distance between the publications with the common subject of research is large. That is why there is a need to use other methods for calculating the distance between texts and estimation of their similarity.

We will improve the second stage of the proposed method for identification of areas of research. To do this, we will generalize the method for finding the distance between publications based on determining the degree of similarity between abstracts. We will assume that the text of abstract S_i, $i=\overline{1,m}$, which was preliminarily processed, corresponds to each article p_i. In particular, each abstract may be presented as a sequence of words in a canonized form after deletion of stop-words [34]. To write down formulas in abstracts, it is proposed to use the way of writing them with the use of TeX [35]. Then the distance between publications may be determined as a degree of similarity of their abstracts. Thus, $g(p_\sigma, p_\tau)=H(S_\sigma, S_\tau)$, where S_σ and S_τ are the abstracts of publications p_σ and p_τ, $\sigma\in\{1,2,\ldots,m\}$, $\tau\in\{1,2,\ldots,m\}$, g is the function of the distance between publications, H is the degree of similarity of publications.

To determine the degree of similarity of abstract H, it is proposed to use the approaches of the n-gram analysis. The text of abstracts S_σ and S_τ in this case is treated as a sequence of n-grams. The task is to find the distance between the abstracts by comparing these n-grams.

Let the assigned abstract S be a fragment of a text. We will consider it as a totality of words. A word is a sequence of characters that belong to finite alphabet $\bar{A}$. Denote the words through $W_n^\beta, W_n^\beta\in S, n\in N$ – a word's number by order, β is the length of a word. Then an arbitrary word is assigned in the form:

$$W_n^\beta=\{t_1,t_2,\ldots,t_\beta\}, \tag{1}$$

where $t_j\in\bar{A}, t_j\notin\bar{C}, j=\overline{1,\beta}$, $\bar{C}$ are all non-letter characters.

Assign the list of stop-words and construct the sequences of words of abstract S in the canonized form, that is:

$$S=\{W_1^{\beta_1},W_2^{\beta_2},\ldots,W_u^{\beta_u}\}, \tag{2}$$

where β_j, $j=\overline{1,u}$ are the lengths of words, and u is the number of words. The sequence of words in the canonized form $\{W_a, W_{a+1}, \ldots, W_{a+n-1}\}$ will be called n-gram. Let W_a and W_b be some words, then the frequency of n-gram is $C(W_a, W_{a+1}, \ldots, W_{a+n-2}, W_b\}$. We will name the ratio of the number of n-gram occurrence in the given text to the total number of n-grams in the text. Then for each pair of words W_a and W_b we will calculate the average frequency of n-gram, which begin with and end in corresponding words, from formula:

$$\mu(W_a,W_b)=$$
$$=\frac{1}{2}\underbrace{\sum_{i_1=1}^{u}\cdots\sum_{i_{n-2}=1}^{u}}_{n-2\text{ times}}\left(\begin{matrix}C\left(W_a,W_{i_1},\ldots,W_{i_{n-2}},W_b\right)+\\+C\left(W_b,W_{i_1},\ldots,W_{i_{n-2}},W_a\right)\end{matrix}\right), \tag{3}$$

where W_{i_j}, $j=\overline{1,n-2}$, $i_j=\overline{1,u}$ is the words in the canonized form that are found in the abstract of the text after canonization, u is the number of words. The degree of similarity of words is determined as:

$$sim(W_a,W_b)=$$
$$=\begin{cases}\dfrac{\ln\dfrac{\mu(W_a,W_b)C_{\max}^2}{C(W_a)C(W_b)\min\{C(W_a),C(W_b)\}}}{-2\ln\dfrac{\min\{C(W_a),C(W_b)\}}{C_{\max}}}, & \text{if } \mu(W_a,W_b)>1,\\ \dfrac{\ln 1.01}{-2\ln\dfrac{\min\{C(W_a),C(W_b)\}}{C_{\max}}}, & \text{if } \mu(W_a,W_b)\le 1,\\ 0, & \text{if } \mu(W_a,W_b)=0,\end{cases} \tag{4}$$

where $C_{\max}$ is the maximum frequency of a word, that is $C_{\max}=\max_i\{C(W_i)\}$, the method for determining the similarity between words is described in detail in paper [9].

Then, to find the similarity of the abstracts of publications, we will use the «text similarity» model based «one to one mapping» [37]. Consider this method in more detail. Let us assume that $S_1=\{W_1,W_2,\ldots,W_{u_1}\}$ and $S_2=\{w_1,w_2,\ldots,w_{u_2}\}$ are the abstracts of two publications in the canonized form, and u_1 and u_2 are the number of words in S_1 and S_2, respectively. Without restricting the commonalities, we will consider that $u_1\le u_2$ Then, we withdraw from the abstracts all coincidences of the words at the same places, i. e. the ones that $W_i=w_i$, $i=\overline{1,u_1}$. Let the number of such coincidences be equal to δ, then if $\delta<u_1$, after coincidences withdrawal we obtain $S_1=\{W_1,W_2,\ldots,W_{u_1-\delta}\}$ and $S_2=\{w_1,w_2,\ldots,w_{u_2-\delta}\}$. For obtained words, we plot the semantic similarity matrix.

$$\begin{pmatrix}\alpha_{11}&\cdots&\alpha_{1(u_{2-\delta})}\\ \vdots&\ddots&\vdots\\ \alpha_{(u_{1-\delta})1}&\cdots&\alpha_{(u_{1-\delta})(u_{2-\delta})}\end{pmatrix}=$$
$$=\begin{pmatrix}sim(W_1,w_1)&\cdots&sim(W_1,w_{u_{1-\delta}})\\ \vdots&\ddots&\vdots\\ sim(W_{u_{1-\delta}},w_1)&\cdots&sim(W_{u_{1-\delta}},w_{u_{2-\delta}})\end{pmatrix}. \tag{5}$$

For each row of the matrix, find a set of elements A_i, the value of which is more than the sum of mathematical expectation and standard root mean square deviation for the given row:

$$A_i=\{\alpha_{ij}, i=\overline{1,u_1}, j=\overline{1,u_2}\mid\alpha_{ij}>M_i+\sigma_i\}, \tag{6}$$

where

$$M_i=\frac{1}{u_{2-\delta}}\sum_{j=1}^{u_{2-\delta}}\alpha_{ij},\ \text{a}\ \sigma_i=\sqrt{\frac{1}{u_{2-\delta}}\sum_{j=1}^{u_{2-\delta}}\left(\alpha_{ij}-M_i\right)^2}. \tag{7}$$

Designate through $\overline{A_i}$ the mathematical expectation of the elements of set A_i. To find the similarity of abstracts, use the formula:

$$H(S_\sigma, S_\tau) = \frac{\left(\sum_{j=1}^{u_2-\delta} \overline{A_i} + \delta\right)(u_2+u_2)}{2u_1u_2}. \quad (8)$$

At stage 2, rather close clusters are united based on the function of distances between publications. To do this, the weight center for each is found and the clusters, the distance between the weight centers of which does not exceed the threshold value, are united [13].

At stage 3, each cluster is put into match with a scientific research area. To establish the match, it is possible to use the expert approach, in which the solution is based on the list of publications of a cluster and some additional information about them, such as keywords, the most widely used concepts, etc.

To conduct the identification of research areas of scientists, at stage 4 we will use the information on the publication activity of scientists, taking into account the constructed set of clusters of research areas, to which these publications belong. Solution to problem of identification of research area of scientists belonging to set $A=\{a_1, a_2, ..., a_n\}$, n is the number of scientists, is mapping $\Lambda: A \to V$, where $V=\{\eta_1, \eta_2, ..., \eta_\psi\}$ is the set of verbal names of scientific research areas, ψ is the number of areas of research. The obtained results allow solving the affine problem of finding the scientists carrying out research in the specified area. The identified areas of research are also the source information for the experts who form a project group, taking into account the experience of each of the potential partners in this or that area. Clustering results give the required amount of information for the solution of the problem of selecting partners, for example, based of fuzzy inference.

5. Construction of fuzzy method of selecting partners for implementation of a scientific research project based on results of clustering

Let a certain project be at the stage of planning, that is, the stages of determining the project environment and of project statement have already been completed. Thus, internal and external factors have been determined, project goals and objectives have been set. A project can be regarded as a totality of separate processes, which involve solving specific problems of a project. Each of the processes requires resources for its implementation. One of the main resources is executives of a task.

Consider the problem of selection of project executives in more detail. To simplify, we will assume that each process from the beginning to the end is performed by one executor. Carry out decomposition of the problem of selecting project executives into sub-problems of selecting one executive for the implementation of each process. At the stage of determining the environment, we formed a set of people who can be executives $A=\{a_1, a_2, ..., a_n\}$, where n is the number of people who can be executors of the process (we will further call them candidates). Each of the candidates can be rated based on the list of criteria $c_1, c_2, ..., c_k$, k is the number of criteria for candidates evaluation.

The aim of the research is to construct the method for evaluation and selection of the rational executor among candidates $a^* \in \{a_1, a_2, ..., a_n\}$ taking into account the set of criteria, at that, the result of using the method can be both one rational executor a^*, and the ordered set of executors $\{a_1^*, a_2^*, \dots a_n^*\}$. The solution in the form of the ordered set of the candidates has a number of advantages. In particular, if the most rational candidate refused to take part in the project due to certain unforeseen factors, the next candidate can be involved in the project.

One of the approaches to the assessment of candidates is to use the multi-criteria group expert estimation. To find generalized aggregated estimates of applicants, it is advisable to apply a system of fuzzy inference. The procedure of fuzzy inference is to determine the mapping of the input data evaluation vector into the scalar output value using fuzzy rules.

Fuzzy inference consists of the following stages:

1. Fuzzification. It is based on some linguistic variables with corresponding linguistic scales. The main procedure of the stage is to determine the degree of belonging of an input value to each of the linguistic variables.

We will consider the mapping obtained as a result of solving the problem of identification of research areas of scientists as discrete fuzzy mapping, which is determined by the ratio of the number of publications of an author in this scientific area to the total number of publications. That is, $\Lambda(a_i) = (\eta_b \,|\, \mu_b^i)$, $b = \overline{1, \psi}$, $i = \overline{1, n}$, and membership is determined from formula:

$$\mu_b^i = \frac{\left\| P(a_i) \cap Y_b \right\|}{\left\| P(a_i) \right\|}, \quad (9)$$

where $P(a_i)$ is the set of all publications of scientist a_i, and Y_b is the cluster of publications that matches research area η_b.

2. Inference mechanism. This stage is based on the base of fuzzy rules, which assign mapping of input fuzzy sets in an output fuzzy set. These rules are formed based of the corresponding expert estimates. There are different procedures for obtaining fuzzy inference, specifically, the procedures of Mamdani, Sugeno, Larsen.

Fuzzy rules are stated in the format «If a candidate has competence η_b with membership degree μ_b, a scientist meets the requirements of a project with membership degree α. For example, the following rule can be stated: «If a candidate knows the methods of project management, the scientist is a likely candidate for the position». In this example, knowing the methods of project management – competence η_b, «excellent» is the verbal qualitative estimate. To transfer qualitative estimates into quantitative values of membership function, it is necessary to use a certain scale. To construct the scale of transition from high-quality expert estimates to membership functions in statements of rules, use the method described in [38]. An example of such a scale is shown in Table 1.

Table 1

Scale for assessment of a statement

No.	Verbal qualitative estimate	Value of membership function
1	Excellent	0.9
2	Good	0.75
3	Satisfactory	0.6
4	Bad	0.35

The inference procedure is the aggregation of all the rules. The process of fuzzy inference in the system «fuzzy inference» is to find degrees of implementation of each of the rules based of the degree of truth of its premise with the help of composition $\alpha=\min\{\mu_b\}$. Fuzzy inference by Mamdani uses the operator of minimum in the system of fuzzy inference by Larsen based on the operator of product.

3. Defuzzification. It involves transformation of a fuzzy value into the clear value. The most common method for defuzzification is the method for finding the weight center of a fuzzy set. Defuzzification takes place with the use of the formula of finding the center of masses.

$$R=\frac{\int_{x_{\min}}^{x_{\max}} x\mu(x)\mathrm{d}x}{\int_{x_{\min}}^{x_{\max}} \mu(x)\mathrm{d}x}, \qquad (10)$$

where x in the fuzzy magnitude, $x_{\min}$ and $x_{\max}$ are the least and the largest values, and $\mu(x)$ is the membership function of fuzzy magnitude. Since in the proposed method, the fuzzy magnitude is discreet, integral should be treated as the sum, that is defuzzification occurs by formula:

$$R^i=\frac{\sum_{b=1}^{\Psi}\alpha^i\mu_b^i}{\sum_{b=1}^{\Psi}\mu_b^i}, \qquad (11)$$

where R^i is the clear estimate of candidate a_i.

The selection of an executor of a scientific project involves finding the executor matching the maximum value of defuzzification R^i. The ordered set of executors is constructed from the scientists at descending defuzzification value.

6. Verification of the method for publications clustering and the fuzzy method for selecting scientific partners

To verify the research, the identification of scientific research areas of 6,0191 scientists that form the knowledge base of the information-analytical system «Base of scientists of Ukraine» was performed. At stage 1, 221,893 publications were clustered using the Louvain method. After discarding 26,495 isolated vertices, 3,328 clusters were obtained. Given the above constraints, 7,044 publications, for which the abstracts were available in English, were separated for subsequent analysis. These publications are split into 281 clusters. Then, to establish the similarities of abstracts (4) to (8) within and between the clusters, the methods of locally sensitive hashing [33] and of n-gram analysis (at $n=3$) were subsequently used. Analysis showed that average similarity of publications within a cluster exceeds by more than 3 times the average similarity of publications of various clusters. That is why the threshold value of 0.09 for uniting was selected. The application of the proposed methods made it possible to reduce the number of clusters to 149 and 134, respectively. Maximum and average values of similarity for both methods have close values and similar behavior (Table 2). The minimum value of similarity within the cluster is higher than the average value for the n-gram analysis. This feature allows identifying the relation between some publications that are distant from the point of view of the method of locally sensitive hashing.

Table 2

Comparison of the methods for finding the distance between publications

Method	Similarity of publications within one cluster			Similarity of publications of different clusters			Number of united clusters
	Min.	Average	Max.	Min.	Average	Max.	
Locally sensitive hashing	0.0025	0.1822	0.3257	0.0023	0.0526	0.1227	132
n-gram analysis ($n=3$)	0.0624	0.1941	0.4125	0.0020	0.0577	0.1128	147

To verify the fuzzy method of scientific partners, it was proposed to determine the organizations-executors for the fundamental scientific research «Methodological foundations for the creation of information environment of management of scientific research of structural units of HEI of MES of Ukraine». As a result, a list of potential organizations-executors that were interested in the project implementation was formed. The executors among professors and teaching staff were selected from each organization. Publication activity was analyzed for each of them. In total, 60,191 Ukrainian scientists were considered. The results of identification of the scientific research areas with the use of the Louvain method at the first stage and results of determining the similarities of abstracts by the method of locally sensitive hash at the second stage were used as well. Subsequently, the requirements for executors of the research according to academic positions of the chief researcher, the senior researcher, etc. were stated. The selection criterion was developed: applicants must have experience in performing studies in the areas «Project management», «Scientometrics» and «Information systems».

Based on (9), degrees of belonging to corresponding competencies that are necessary for the implementation of various stages of fundamental research were designed. The obtained result determined a ranking list of scientists engaged in the study of theoretical aspects of the scientific research activity. Using the fuzzy method for identification of research project groups based on expert estimation, it was found that two universities can be executors of the project: Kyiv National University of Building and Architecture, as well as Odessa National Polytechnic University.

7. Discussion of results of improvement of the method for publications clustering based on n-gram analysis and fuzzy method of selecting research partners

It was shown in this study that the graph clustering requires additional information that can be provided by the measure of abstracts proximity based on n-gram analysis. Availability of the additional information allows reducing the number of clusters by several times. The problem part in the fuzzy method for identifying research partners is that it is necessary to make a choice of experts of a narrow profile. Due to the modified method of clustering and statement of requirements for potential partners, this problem is partially solved.

Most research into selecting partners usually involve experts or conduct cross-sectional evaluation of partners

between one another, in this case, the project management apparatus and methods of processing expert information are used to great extent. This research provides an opportunity to calculate the merits of potential partners without any influence of the subjective factor due to the application of the method for determining scientific research areas.

For correct operation of the proposed method for calculating the distance between text data, it is important to use true information about the frequency of *n*-gram. Google Ngram Viewer [39] is the search online service of the Google company that allows plotting the graphs of frequency of language units based on a huge number of printed sources. Statistic information is available for the American and British variants of English, French, German, Spanish, Italian, Russian and simplified Chinese languages and Hebrew. Similar information for the Russian language is also provided by the National corpus of Russian language [40]. For the Ukrainian language, information on monograms and bigrams is available in the corpus of the Ukrainian language in linguistic portal Mova [41]. The volumes of available data are shown in Table 3. In paper [42], it is noted that about 6 % of all textual information generated by humanity throughout the entire history has been processed until now.

Table 3

Volumes of *n*-gram frequencies

Source	Number of texts	Number of *n*-gram
Google Ngram Viewer – English	4,541,627	468,491,999,592
Google Ngram Viewer – Russian	591,310	67,137,666,353
National corpus of the Russian language	76,882	209,198,275
Corpus of the Ukrainian language	Data unavailable	87,180,005

Thus, we can conclude that the volumes of data of relative frequencies of *n*-gram occurrence for the Ukrainian language are considerably smaller than for the English language. This can become a significant constraint of the proposed method. To overcome it, it is proposed to use the abstracts in English.

Based on the research results, the fuzzy method for choosing the partners for the implementation of a scientific research project was proposed, but it was not determined how the contribution of selected partners to a project will be estimated. This is the issue for future research. One of the ways of research development is to create new methods for identification of similarities between the fragments of text information. The research development is seen in the creation of a possibility for dynamic management of partners selection.

8. Conclusions

1. The two-stage method for clustering the publications of scientists was further developed in this research. At stage 1, the publications citation graph is clustered. At stage 2, clusters are united based on the criterion of proximity of abstracts of publications. The improvement of the method implies the use of the method of *n*-gram analysis for calculating the distance between publications, which increases the accuracy of determining the distance between the text information. Compared with the locally-sensitive hashing, the application of the method of *n*-gram analysis allows establishing a relation between some publications that are distant from the point of view of the method of locally-sensitive hashing. At the same time, a significant constraint of using this method is insufficient volume of data about the relative frequency of *n*-grams for the Ukrainian language.

2. The new method for choosing partners for the implementation of a scientific research project was designed. Its essence is the use of the fuzzy inference to coordinate the opinions of experts on the necessary requirements for the candidates. Results of clustering of scientists» publications are applied for fuzzification of scientists matching the research tasks of a project. For defuzzification of the estimates of scientists that is based on the requirements, stated by experts, the method of finding the weight center was applied. The merit of the method is the possibility of statement of requirements in the linguistic form.

3. To verify the proposed methods, publications clustering was performed taking into account the similarity of abstract based on the *n*-gram analysis. The selection of executors for fundamental scientific research was performed based on the results of clustering.

References

1. Šubelj, L., van Eck, N. J., Waltman, L. (2016). Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods. PLOS ONE, 11 (4), e0154404. doi: https://doi.org/10.1371/journal.pone.0154404
2. Dhillon, I. S., Guan, Y., Kulis, B. (2007). Weighted Graph Cuts without Eigenvectors A Multilevel Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (11), 1944–1957. doi: https://doi.org/10.1109/tpami.2007.1115
3. Waltman, L., van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. The European Physical Journal B, 86 (11). doi: https://doi.org/10.1140/epjb/e2013-40829-0
4. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008 (10), P10008. doi: https://doi.org/10.1088/1742-5468/2008/10/p10008
5. Yang, J., Leskovec, J. (2013). Overlapping community detection at scale. Proceedings of the Sixth ACM International Conference on Web Search and Data Mining - WSDM «13, 587–596. doi: https://doi.org/10.1145/2433396.2433471
6. Pons, P., Latapy, M. (2006). Computing Communities in Large Networks Using Random Walks. Journal of Graph Algorithms and Applications, 10 (2), 191–218. doi: https://doi.org/10.7155/jgaa.00124
7. Bolelli, L., Ertekin, S., Giles, C. L. (2006). Clustering Scientific Literature Using Sparse Citation Graph Analysis. Knowledge Discovery in Databases: PKDD 2006, 30–41. doi: https://doi.org/10.1007/11871637_8
8. Gomaa, W. H., Fahmy, A. A. (2013). A Survey of Text Similarity Approaches. International Journal of Computer Applications, 68 (13), 13–18. doi: https://doi.org/10.5120/11638-7118
9. Islam, A., Milios, E., Kešelj, V. (2012). Text Similarity Using Google Tri-grams. Lecture Notes in Computer Science, 312–317. doi: https://doi.org/10.1007/978-3-642-30353-1_29

10. Brants, T., Franz, A. (2006). Web 1T 5-gram corpus version 1.1. Technical report. Google Research.
11. Kuchansky, A., Andrashko, Y., Biloshchytskyi, A., Danchenko, E., Ilarionov, O., Vatskel, I., Honcharenko, T. (2018). The method for evaluation of educational environment subjects» performance based on the calculation of volumes of msimplexes. Eastern-European Journal of Enterprise Technologies, 2 (4 (92)), 15–25. doi: https://doi.org/10.15587/1729-4061.2018.126287
12. Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S., Kuzka, O., Terentyev, O. (2017). Evaluation methods of the results of scientific research activity of scientists based on the analysis of publication citations. Eastern-European Journal of Enterprise Technologies, 3 (2 (87)), 4–10. doi: https://doi.org/10.15587/1729-4061.2017.103651
13. Teslia, I., Latysheva, T. (2016). Development of conceptual frameworks of matrix management of project and programme portfolios. Eastern-European Journal of Enterprise Technologies, 1 (3 (79)), 12–18. doi: https://doi.org/10.15587/1729-4061.2016.61153
14. Yazici, H. J. (2009). The Role of Project Management Maturity and Organizational Culture in Perceived Performance. Project Management Journal, 40 (3), 14–33. doi: https://doi.org/10.1002/pmj.20121
15. Morozov, V., Kalnichenko, O., Liubyma, I. (2017). Managing projects configuration in development distributed information systems. 2017 2nd International Conference on Advanced Information and Communication Technologies (AICT). doi: https://doi.org/10.1109/aiact.2017.8020088
16. Su, Z., Poulin, D. (1996). Partnership management within the virtual enterprise in a network. IEMC 96 Proceedings. International Conference on Engineering and Technology Management. Managing Virtual Enterprises: A Convergence of Communications, Computing, and Energy Technologies. doi: https://doi.org/10.1109/iemc.1996.547894
17. Talluri, S., Baker, R. C. (1996). A quantitative framework for designing efficient business process alliances. IEMC 96 Proceedings. International Conference on Engineering and Technology Management. Managing Virtual Enterprises: A Convergence of Communications, Computing, and Energy Technologies. doi: https://doi.org/10.1109/iemc.1996.547896
18. XueNing, C., Tso, S. K., Zhang, W. J., Li, Q. (2000). Partners selection for virtual enterprises. Proceedings of the 3rd World Congress on Intelligent Control and Automation (Cat. No.00EX393). doi: https://doi.org/10.1109/wcica.2000.859940
19. Feng, W. D., Chen, J., Zhao, C. J. (2000). Partners selection process and optimization model for virtual corporations based on genetic algorithms. Journal of Tsinghua University (Science and Technology), 40, 120–124.
20. Biloshchytskyi, A., Biloshchytska, S., Kuchansky, A., Bielova, O., Andrashko, Y. (2018). Infocommunication system of scientific activity management on the basis of project-vector methodology. 2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET). doi: https://doi.org/10.1109/tcset.2018.8336186
21. Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S., Danchenko, O. (2018). Development of Infocommunication System for Scientific Activity Administration of Educational Environment's Subjects. 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T). doi: https://doi.org/10.1109/infocommst.2018.8632036
22. Biloshchytskyi, A., Kuchansky, A., Paliy, S., Biloshchytska, S., Bronin, S., Andrashko, Y. et. al. (2018). Development of technical component of the methodology for projectvector management of educational environments. Eastern-European Journal of Enterprise Technologies, 2 (2 (92)), 4–13. doi: https://doi.org/10.15587/1729-4061.2018.126301
23. Mulesa, O., Geche, F. (2016). Designing fuzzy expert methods of numeric evaluation of an object for the problems of forecasting. Eastern-European Journal of Enterprise Technologies, 3 (4 (81)), 37–43. doi: https://doi.org/10.15587/1729-4061.2016.70515
24. Kuchansky, A., Biloshchytskyi, A. (2015). Selective pattern matching method for time-series forecasting. Eastern-European Journal of Enterprise Technologies, 6 (4 (78)), 13–18. doi: https://doi.org/10.15587/1729-4061.2015.54812
25. Kuchansky, A., Biloshchytskyi, A., Andrashko, Y., Biloshchytska, S., Shabala, Y., Myronov, O. (2018). Development of adaptive combined models for predicting time series based on similarity identification. Eastern-European Journal of Enterprise Technologies, 1 (4 (91)), 32–42. doi: https://doi.org/10.15587/1729-4061.2018.121620
26. Mulesa, O., Geche, F., Batyuk, A., Buchok, V. (2017). Development of Combined Information Technology for Time Series Prediction. Advances in Intelligent Systems and Computing, 361–373. doi: https://doi.org/10.1007/978-3-319-70581-1_26
27. Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S., Dubnytska, A., Vatskel, V. (2017). The method of the scientific directions potential forecasting in infocommunication systems of an assessment of the research activity results. 2017 4th International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T). doi: https://doi.org/10.1109/infocommst.2017.8246352
28. Snytyuk, V. E. (2008). Forecasting. Models. Methods. Algorithms. Kyiv: Maklaut, 364.
29. Sihombing, D. I., Sitompul, O. S., Sutarman, Nababan, E. (2018). Combining the use of analytical hierarchy process and lexicographic goal programming in selecting project executor. IOP Conference Series: Materials Science and Engineering, 420, 012113. doi: https://doi.org/10.1088/1757-899x/420/1/012113
30. Asanov, A., Myshkina, I. (2017). Selection of executors for realization of individual tasks of the project. SHS Web of Conferences, 35, 01026. doi: https://doi.org/10.1051/shsconf/20173501026
31. Imangulova, Z., Kolesnyk, L. (2016). An algorithm for building a project team considering interpersonal relations of employees. Eastern-European Journal of Enterprise Technologies, 6 (3 (84)), 19–25. doi: https://doi.org/10.15587/1729-4061.2016.85222
32. Baiden, B. K., Price, A. D. F. (2011). The effect of integration on project delivery team effectiveness. International Journal of Project Management, 29 (2), 129–136. doi: https://doi.org/10.1016/j.ijproman.2010.01.016
33. Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S., Kuzka, O., Shabala, Y., Lyashchenko, T. (2017). A method for the identification of scientists» research areas based on a cluster analysis of scientific publications. Eastern-European Journal of Enterprise Technologies, 5 (2 (89)), 4–11. doi: https://doi.org/10.15587/1729-4061.2017.112323

34. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Biloshchytska, S., Chala, L. (2016). Detection of near dublicates in tables based on the locality-sensitive hashing method and the nearest neighbor method. Eastern-European Journal of Enterprise Technologies, 6 (4 (84)), 4–10. doi: https://doi.org/10.15587/1729-4061.2016.86243
35. Biloshchytskyi, A., Kuchansky, A., Biloshchytska, S., Dubnytska, A. (2017). Conceptual model of automatic system of near duplicates detection in electronic documents. 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM). doi: https://doi.org/10.1109/cadsm.2017.7916155
36. Trzeciak, J. (2005). Writing Mathematical Papers in English. A Practical Guide. European Mathematical Society, 51. doi: https://doi.org/10.4171/014
37. Islam, A., Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data, 2 (2), 1–25. doi: https://doi.org/10.1145/1376815.1376819
38. Biloshchytskyi, A., Myronov, O., Reznik, R., Kuchansky, A., Andrashko, Y., Paliy, S., Biloshchytska, S. (2017). A method to evaluate the scientific activity quality of HEIs based on a scientometric subjects presentation model. Eastern-European Journal of Enterprise Technologies, 6 (2 (90)), 16–22. doi: https://doi.org/10.15587/1729-4061.2017.118377
39. Ngram Viewer. Available at: https://books.google.com/ngrams
40. National corpus of Russian language. Available at: http://www.ruscorpora.ru/new/index.html
41. National corpus of Ukrainian language. Available at: http://www.mova.info/corpus.aspx
42. Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 169–174.

Система функцій Фабера-Шаудера була введена в 1910 році і стала першим прикладом базису в просторі функцій, неперервних на [0, 1]. Відомо низку результатів щодо властивостей рядів Фабера-Шаудера, у тому числі щодо оцінювання похибок наближення функцій поліномами та частинними сумами рядів, побудованих за системою Фабера-Шаудера. Відомо, що серед завдань теорії наближення функцій важливим є отримання нових оцінок величини наближення довільної функції деяким заданим класом функцій. Тому дослідження апроксимативних властивостей поліномів і частинних сум рядів Фабера-Шаудера становить значний інтерес для сучасної теорії апроксимації функцій.

Досліджено питання наближення функцій обмеженої варіації частинними сумами рядів, побудованих за системою функцій Фабера-Шаудера. Отримано оцінку похибки апроксимації функцій з класів функцій обмеженої варіації C_p ($1\le p<\infty$) у метриці простору L_p за допомогою значень модуля неперервності дробового порядку $\varpi_{2-1/p}(f, t)$. З отриманої нерівності випливає оцінка похибки наближення неперервних функцій, яка виражена через модуль неперервності другого порядку.

Також у класі функцій C_p ($1<p<\infty$) отримані оцінки похибок наближення функцій у метриці простору L_p за допомогою модуля неперервності дробового порядку $\varpi_{1-1/p}(f, t)$.

Для класів функцій обмеженої варіації $KCV_{(2,p)}$ ($1\le p<\infty$) отримано оцінку похибки наближення функцій у метриці простору L_p частинними сумами рядів Фабера-Шаудера.

Таким чином, отримано низку оцінок похибок наближення функцій обмеженої варіації їх частинними сумами рядів Фабера-Шаудера. Отримані результати є новими у теорії наближення функцій. Вони певним чином узагальнюють раніше відомі результати та можуть бути використані для подальших практичних застосувань.

Ключові слова: функції обмеженої варіації, інтегральна метрика, модуль неперервності, система Фабера-Шаудера.

UDC 517.5

DOI: 10.15587/1729-4061.2019.176595

A STUDY OF APPROXIMATION OF FUNCTIONS OF BOUNDED VARIATION BY FABER-SCHAUDER PARTIAL SUMS

N. Mormul
PhD, Associate Professor
Department of Mathematics and Information
University of Customs and Finance
Vladimir Vernadsky str., 2/4, Dnipro, Ukraine, 49000
E-mail: nikolaj.mormul@gmail.com

A. Shchitov
PhD, Associate Professor
Independent Researcher
E-mail: an_shchitov@rambler.ru

Received date 01.07.2019
Accepted date 05.08.2019
Published date 29.08.2019

1. Introduction

The Faber-Schauder system of functions was introduced in the paper [1] and became the first example of a basis of the space of functions continuous on [0, 1]. Approximate properties of the Faber-Schauder system regarding approximation of individual functions and classes of functions are studied, for example, in [2–5]. In those studies, the upper bounds of