Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 16, NOVEMBER 2023

DOI: 10.37943/16AADE3851

Andrii Biloshchytskyi
Doctor of Technical Sciences, Professor, Vice-Rector for Science and Innovation
a.b@astanait.edu.kz, orcid.org/0000-0001-9548-1959
Astana IT University, Kazakhstan
Professor of the Department of Information Technologies,
Kyiv National University of Construction and Architecture, Ukraine

Olexandr Kuchansky*
Doctor of Technical Sciences, Head of the Department of Information Systems and Technologies
kuczanski@gmail.com, orcid.org/0000-0003-1277-8031
Taras Shevchenko National University of Kyiv, Ukraine
Visiting Professor at Astana IT University, Kazakhstan

Aidos Mukhatayev
Candidate of Pedagogical Sciences, Associate Professor, Director Bologna Process and academic mobility center
mukhatayev.aidos@gmail.com, orcid.org/0000-0002-8667-3200
Chief Researcher Astana IT University, Kazakhstan

Svitlana Biloshchytska
Doctor of Technical Sciences, Associate Professor, Professor of the Department of Computational and Data Science
bsvetlana2007@gmail.com, orcid.org/0000-0002-0856-5474
Astana IT University, Kazakhstan
Professor of the Department of Information Technologies,
Kyiv National University of Construction and Architecture, Ukraine

Yurii Andrashko
PhD, Associate Professor, Department of System Analysis and Optimization Theory
yurii.andrashko@uzhnu.edu.ua, orcid.org/0000-0003-2306-8377
Uzhhorod National University, Ukraine

Sapar Toxanov
PhD candidate
sapar6@mail.ru, orcid.org/0000-0002-2915-9619
D. Serikbayev East Kazakhstan Technical University, Kazakhstan

Adil Faizullin
PhD candidate
adil.faizullin@astanait.edu.kz, orcid.org/0000-0001-5644-9841
Manash Kozybayev North Kazakhstan University, Kazakhstan

CLUSTERING OF SCIENTISTS' PUBLICATIONS, CONSIDERING FINDING SIMILARITIES IN ABSTRACTS AND TEXTS OF PUBLICATIONS BASED ON N-GRAM ANALYSIS AND IDENTIFYING POTENTIAL PROJECT GROUPS

Abstract: The article describes the solution to the problem of clustering scientists' publications, taking into account the finding of similarities in the annotations and texts of these publications based on n-grams of analysis and cross-references, as well as the tasks of identifying potential project groups for the implementation of research and educational projects

Received: 27.09.2023 Accepted 26.10.2023 Published 14.11.2023

DOI: 10.37943/16AADE3851

based on the results of clustering. The selection of scientific partners in the world practice is done without a comprehensive assessment of their activities. Most of the well-known indexes for evaluating the research activities of scientists need to consider information about citations fully. The methods developed in the study for evaluating the scientific activities of scientists and universities, as well as methods for selecting scientific partners for the implementation of educational and scientific projects on a scientific basis, allow us to organize the influential work of universities qualitatively. In the article, a probabilistic thematic model is constructed that allows the clustering of scientists' publications in scientific fields, considering the citation network, which is an important step in solving the problem of identifying subject scientific spaces. As a result of constructing the model, the problem of increasing instability of clustering of the citation graph due to a decrease in the number of clusters has been solved. The main objective of this work is to address the challenge of selecting suitable partners for collaboration in scientific and educational projects. To achieve this, a method for choosing project executors has been developed, which employs fuzzy logical inference to harmonize expert opinions regarding candidate requirements. This approach helps facilitate the multi-criteria selection of potential partners for scientific and educational projects. In addition to the method, various software modules have been created as part of this research. These modules are designed for the automated collection of information on the publications and citation records of scientists through international scientometric databases. They also encompass a visualization module and a user interface that aids in evaluating the scientific activities of university teaching staff. Choosing partners for grants or strategic collaborations, especially in the context of a globalized and highly mobile scientific community, remains a pertinent issue. The approach described in this research involves clustering the scientific publications of potential project partners. Furthermore, it incorporates conducting comparative citation analyses of these publications and establishing proximity based on n-gram annotation analysis. These methods provide a scientific basis for making informed choices when selecting partners, which is crucial for initiating and advancing research projects. Consequently, the selection of partners for forming research project teams is an immediate and pressing task.

Keywords: scientometry; search for scientific partners; scientific collaboration; clustering of publications; n-gram analysis; determination of research directions.

Introduction

The ongoing progress of a country's scientific landscape plays a pivotal role in enhancing its prestige, driving economic growth, and fostering innovation across various realms of human endeavor. In recent decades, researchers have dedicated themselves to a crucial mission: devising mechanisms for adeptly managing the evolution of the scientific landscape. This involves drawing in private entities, securing financial backing from governmental bodies at different tiers, and fostering international collaboration through specific scientific and educational initiatives like Horizon 2020 and Erasmus+. Consequently, a vital objective for both the state and private enterprises keen on advancing high-tech technologies and engaging foreign partners is the establishment of effective criteria for evaluating the outcomes of research endeavors undertaken by scientists, higher education institutions, and their respective units.

In order to modernize the social development of the Republic of Kazakhstan in the context of global globalization processes, it is imperative to increase the level of research and education in general consistently. The successful solution of Kazakhstan's problems on its integration into the world economy is only possible with highly qualified specialists. The changes necessary for this in the education system of Kazakhstan consist of organizational, technological, and functional improvements that will allow us to reach the level of leading European educational institutions.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 16, NOVEMBER 2023

International experience shows that the development and implementation of national systems for evaluating the performance of a university has a positive effect on the effectiveness of its functioning. The existing system of evaluating the performance of universities by the central education authorities does not aim for the university to introduce new principles and mechanisms for achieving and continuously improving the quality of scientific and educational activities but only fixes specific and not always reliable indicators without analyzing the possible reasons that contributed to their formation. The selection of scientific partners in the world practice is done without a comprehensive assessment of their activities. Most of the well-known indexes for evaluating the research activities of scientists need to take into account information about citations fully. The methods developed in the study for evaluating the scientific activities of scientists and universities and methods for selecting scientific partners to implement educational and scientific projects on a scientific basis allow us to organize the influential work of universities qualitatively.

Evaluating the outcomes of research activities serves the purpose of verifying whether the research process aligns with the initially defined goals and allows for adjustments if needed. Scientific publications stand as a key component in assessing research work, with their significance often gauged by the extent to which their results are utilized in other research endeavors.

The holistic evaluation of research can rely on subjective judgments from scientists themselves. However, widely accepted criteria for appraising the outcomes of scientific research involve the citation indicators of publications, typically represented as scalar values. While this approach offers several advantages, it is not without drawbacks. These include potential data loss and instances where the parameter fails to change with an increase in citations and publications. Consequently, there is a need to develop new methods or modify existing ones to evaluate the results of scientific research activities, addressing these limitations. In contemporary settings, the challenge of enhancing the efficiency of scientific research and fostering practical collaboration within scientific communities has become increasingly pressing. The creation of teams that can implement projects successfully is of particular importance for project-oriented organizations. To form the team for a research project, a typical approach is to select partners from among scientists who have the appropriate qualifications and experience in implementing such projects.

Selecting partners for scientific research projects, especially in the context of globalization and increased mobility of scientific communities, is crucial. One effective approach to address this challenge involves clustering the scientific publications of potential project partners. Additionally, conducting a comparative citation analysis and establishing proximity between these publications through n-gram annotation analysis is essential for making informed and rational choices. These methods provide scientific validation for partner selection, laying the foundation for the creation and advancement of research projects in the future. Therefore, the process of choosing partners for research projects is a timely and important undertaking.

Overview of publications in this field

A crucial metric for assessing the performance of higher education institutions involves the quantity of scientific publications authored by university staff and indexed in scientometric databases, along with citation indicators reflecting the impact of a university's scientific publications and their qualitative aspects. Classical citation indices, such as the h-index and g-index, are commonly employed to calculate these indicators, as discussed in [1, 2].

The transition from qualitative assessments to quantitative ones is explored in [3, 4], but this shift is not without drawbacks, particularly the reliance on expert assessments, introducing a subjective element into the evaluation process. In [5], the ideal point method is

DOI: 10.37943/16AADE3851

employed to address adaptive selection problems. While [6, 7] delve into expert methods for evaluating research activities, they may not comprehensively cover crucial components such as the citation of scientists' publications in scientometric databases. In [8], a model for evaluating the activities of higher education institutions is considered, which allows for a transition from university assessments to forecasting the prospects for their development in the future. The paper [9] describes a method for predicting the potential of research directions at universities. Indeed, a comprehensive assessment of a higher education institution's potential necessitates considering its multifaceted activities, including scientific, educational, international, innovative, and various other dimensions. The evaluation process should encompass a holistic view that takes into account the institution's diverse contributions and impact across different domains. This ensures a more accurate and nuanced understanding of the institution's overall potential and effectiveness in fulfilling its mission.

One of the components of the evaluation of the activities of universities around the world is the determination of a generalized indicator of the quality and results of scientific research of an individual scientist, department, faculty, and university. In the modern world of information technology, many publications that are available in the web space allow you to assess the scientific level of research. However, the lack of uniform requirements and standards for the placement and management of scientific works creates natural obstacles to the qualitative assessment of the results of the activities of scientometric subjects. Solving this problem requires [10]:

1. Definition of the leading entities of the subjects of scientometry and the links between them;

2. Creating an appropriate degree of formalization of the management processes of scientific publications at different stages of their processing;

3. Establishing a worldwide database encompassing scientists, scientific publications, journals, and institutions with the aim of assessing citation ratings and gauging the popularity of these entities.

The evaluation of scientists' research activities often involves assessing the citation indicators of their published works. In [2, 11], there is a comprehensive exploration of scientometric databases and methods for acquiring prominent citation indicators. Among these, the Hirsch index remains the most prevalent bibliometric indicator to date, as detailed in [1]. The Hirsch index is determined by the number of articles (h) a scientist has published, each of which must be cited at least h times.

In [12], an alternative index, the g-index, is proposed. This index is represented by a value (g) corresponding to the number of articles cited at least g^2 times in total. However, [13, 14] points out fundamental drawbacks of both the h- and g-indexes, notably the loss of information regarding the citation of an author's most popular publications. To address these limitations, the e-index is suggested as a remedy in [13, 14].

The primary drawback is that each of the mentioned indexes lacks some citation information:

- The h-index disregards information beyond the Hirsch core (h-core): publications cited fewer than h times and citations exceeding h times are not considered.

- The g-index omits information beyond the g-core, depending on the citation ratio and the author's number of publications.

- the e-index loses information about the citation of publications when cited less than h times.

- the I-10 index loses information about the publication when cited less than 10 times.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 16, NOVEMBER 2023

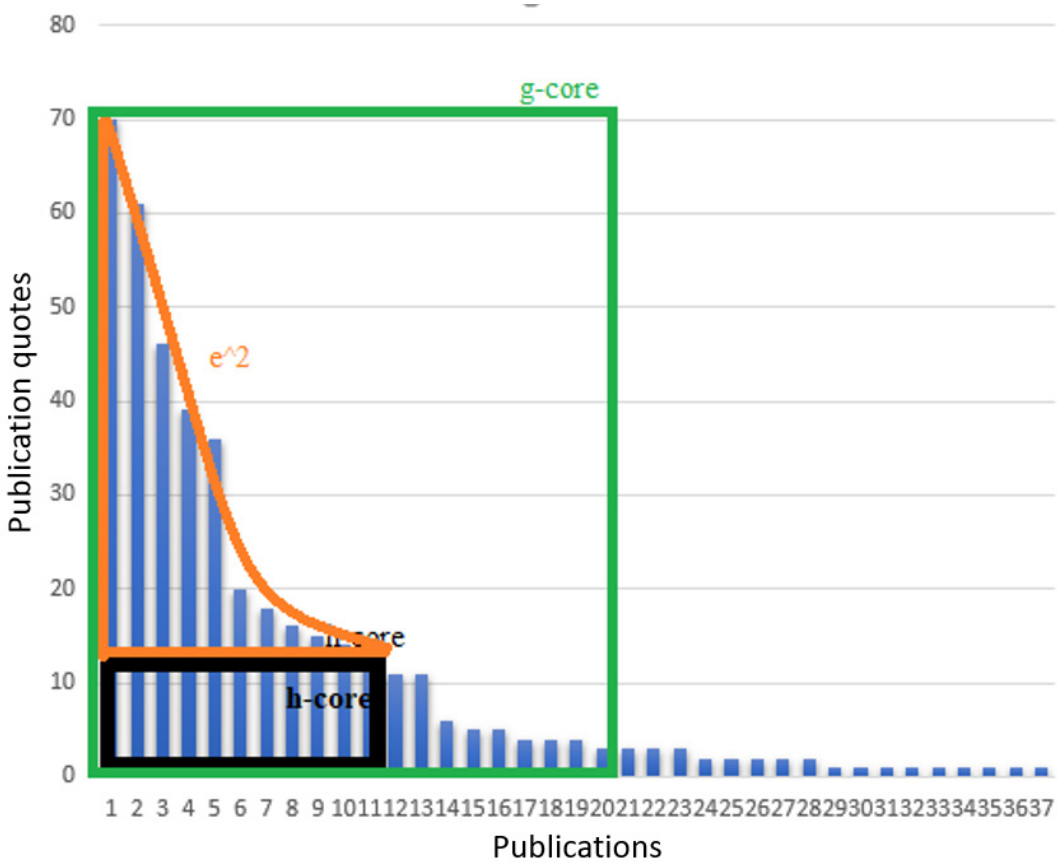

Figure 1. Graphical representation of scientometric index cores

One of the drawbacks associated with well-known citation calculation indices, such as the Hirsch index, the I-10 index, and the g-index, is the existence of limiting cases where these parameters fail to change their values despite an increase in the number of citations and publications. To illustrate this, consider a scenario where a scientist has published fundamental research papers in a specific field, and these publications have become widely influential, with multiple citations each. If the number of these citations, denoted as "di," exceeds a certain threshold (n), traditional bibliometric indexes would become saturated at n'. In such cases, the traditional bibliometric indexes might inaccurately suggest that the scientist's research is neither very successful nor significant. This inadequacy highlights the need for new methods that can accurately reflect the progress of research as new publications emerge, get cited, and thus ensure that the evaluation of research results remains meaningful. It's essential for these new methods to address and rectify this limitation seen in traditional citation indices.

Formulation and description of problems of clustering of scientists' publications, taking into account finding similarities in the annotations and texts of these publications based on n-gram analysis and cross-references.

In [16], the concept of identifying the research directions of scientists is described as the process of linking a particular scientist to the scientific domains in which they operate and publish their research papers.

In [16], a method for identifying research directions of scientists was proposed, consisting of four stages:

1. Clustering of the citation graph of publications.
2. Consolidation of clusters.
3. Building correspondence between clusters of publications.
4. Identification of research directions of scientists.

DOI: 10.37943/16AADE3851

The process begins with considering a set of publications denoted as $P=\{p1,p2,...,pm\}$. These publications serve as vertices in a citation graph, and connections between them are represented by citation arcs. The clustering procedure is then applied to group these scientific publications into distinct clusters, resulting in the formation of a set of clusters denoted as Y. Given that the size of the set Y can be substantial, there arises a need to enhance the constructed clusters by merging those that are in close proximity and have a small number of elements. The distance between publications is determined based on the meaningful proximity of annotations by content, as proposed in the method outlined in reference [16]. This method utilizes locally sensitive hashing to compute distances between annotations. The approach yields satisfactory results, particularly when comparing annotations created by scientists within the same region.

It's important to note that the cultural, regional, and linguistic traditions of authors significantly influence writing styles and the use of specific words and phrases. While the proposed method works well for comparing annotations from scientists in the same region, disparities arise when comparing annotations from scientists in different countries. In such cases, the distance between publications with a shared research subject may be considerable. Therefore, alternative methods are necessary to determine the distance between texts when comparing annotations from diverse cultural and regional backgrounds. A universal method for calculating the distance between publications involves assessing the similarity of their annotations. Let each article pi be associated with a preprocessed textual representation of its annotation S_i, $i = \overline{1,m}$. The abstracts are processed to represent a sequence of words in a standardized form after removing stop words [17]. Formulas in the instructions are represented in their textual form using TeX [18]. The distance between publications is then defined as the similarity degree of their annotations, i.e., $g(p_\sigma, p_\tau) = H(S_\sigma, S_\tau)$, where Sσ and Sτ are the annotations of publications pσ and pτ, $\sigma \in \{1,2,\dots,m\}$, $\tau \in \{1,2,\dots,m\}$, g is the distance function between publications, and H is the similarity degree of annotations.

To determine the similarity degree H, n-gram analysis approaches are proposed. The annotations' text, S_σ and S_τ, are treated as sequences of n-grams, and the task is to find the distance between annotations by comparing these n-grams.

Let's consider an annotation S, which is a text fragment composed of words. Let $\bar{A}$ be a sequence of characters from a finite alphabet. The authors denote the words as $W_n^\beta, W_n^\beta \in S, n \in N$ –represents the ordinal number of the word, and β is the length of the word. An arbitrary word can be expressed as:

$$W_n^\beta = \{t_1, t_2, \dots, t_\beta\}, \quad (1)$$

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

Let's establish a set of stop words and create sequences of annotation words S in a canonicalized form, that is,

$$S = \left\{W_1^{\beta_1}, W_2^{\beta_2}, \dots, W_u^{\beta_u}\right\}, \quad (2)$$

where β_j, $j = \overline{1,u}$ – word lengths, and u is the number of words.

Let us call an n-gram sequence of words in the canonized form $\{W_a, W_{a+1}, \dots, W_{a+n-1}\}$. If Wa and Wb represent certain words, the frequency of an n-gram $C(W_a, W_{a+1}, \dots, W_{a+n-2}, W_b\}$ can be defined as the ratio of the number of occurrences of that n-gram in a specific text to the total number of n-grams in the text. For each pair of words W_a and W_b, the average frequency of n-grams starting and ending with the respective words can be calculated using the formula:

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 16, NOVEMBER 2023

$$\mu(W_a, W_b) = \frac{1}{2}\underbrace{\sum_{i_1=1}^{u} \dots \sum_{i_{n-2}=1}^{u}}_{\substack{n-2 \\ times}} 0\left(C(W_a, W_{i_1}, \dots, W_{i_{n-2}}, W_b) + C(W_b, W_{i_1}, \dots, W_{i_{n-2}}, W_a)\right)$$

where $W_{i_j}, j = \overline{1, n-2}, i_j = \overline{1, u}$,

– The degree of similarity of words, after canonization, in canonized form, which occur in the annotation is determined as:

$$\mathrm{sim}(\mathrm{W_a}, \mathrm{W_b}) = \begin{cases} \dfrac{\ln \dfrac{\mu(\mathrm{W_a}, \mathrm{W_b})\mathrm{C}_{\max}^2}{\mathrm{C(W_a)C(W_b)}\min\{\mathrm{C(W_a)}, \mathrm{C(W_b)}\}}}{-2\ln \dfrac{\min\{\mathrm{C(W_a)}, \mathrm{C(W_b)}\}}{\mathrm{C}_{\max}}} & \text{if} \quad \mu(\mathrm{W_a}, \mathrm{W_b}) > 1 \\ \dfrac{\ln 1.01}{-2\ln \dfrac{\min\{\mathrm{C(W_a)}, \mathrm{C(W_b)}\}}{\mathrm{C}_{\max}}} & \text{if} \quad \mu(\mathrm{W_a}, \mathrm{W_b}) \le 1 \\ 0 & \text{if} \quad \mu(\mathrm{W_a}, \mathrm{W_b}) = 0 \end{cases} \tag{3}$$

where $\mathrm{C}_{\max}$ is the maximum frequency of a word, i.e. $Cmax_{i_{max}}$, the method for determining the similarity between words is described in more detail [19].

Next, to find the similarity of publication annotations, the "text similarity model" based on "one-to-one mapping" is applied [20]. Let us consider this method in more detail. Let $S_1 = \{W_1, W_2, \dots, W_{u_1}\}$ and $S_2 = \{w_1, w_2, \dots, w_{u_2}\}$ the annotations of the two publications are in canonical form, and u1 and u2 are the number of words in S1 and S2, respectively. Without limiting the generality, the authors assume that $\mathrm{u1} \le \mathrm{u2}$, then all the matches of words in the same places are removed from the annotations, that is, such that $\mathrm{W_i}=\mathrm{w_i} i = \overline{1, u_1}$. Let the number of such matches be equal to δ, then if $\delta < \mathrm{u1}$ then after extracting the matches the authors get $S_1 = \{W_1, W_2, \dots, W_{u_1-\delta}\}$ and $S_2 = \{w_1, w_2, \dots, w_{u_2-\delta}\}$. For the obtained sequences of words, a matrix of semantic similarity is constructed.

$$\begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1(u_2-\delta)} \\ \vdots & \ddots & \vdots \\ \alpha_{(u_1-\delta)1} & \cdots & \alpha_{(u_1-\delta)(u_2-\delta)} \end{pmatrix} = \begin{pmatrix} sim(W_1, w_1) & \cdots & sim(W_1, w_{u_1-\delta}) \\ \vdots & \ddots & \vdots \\ sim(W_{u_1-\delta}, w_1) & \cdots & sim(W_{u_1-\delta}, w_{u_2-\delta}) \end{pmatrix} \tag{4}$$

For every row in this matrix, a set of elements Ai whose value exceeds the sum of the mean and the standard deviation for that particular row are identified.

$$A_i = \{\alpha_{ij},\ i = \overline{1, u_1},\ j = \overline{1, u_2} | \alpha_{ij} > M_i + \sigma_i,\} \tag{5}$$

where $M_i = \frac{1}{u_2-\delta}\sum_{j=1}^{u_2-\delta} \alpha_{ij}$, while $\sigma_i = \sqrt{\frac{1}{u_2-\delta}\sum_{j=1}^{u_2-\delta}(\alpha_{ij} - M_i)^2}$.

Denote the mean of the elements in the set $\mathrm{A_i}$ as $\overline{A_i}$. To determine the similarity of annotations, the formula is utilized: (5)

$$H(S_\sigma, S_\tau) = \frac{\left(\sum_{j=1}^{u_2-\delta} \overline{A_i} + \delta\right)(u_2+u_2)}{2u_1u_2} \tag{6}$$

DOI: 10.37943/16AADE3851

In the second phase, clusters that are deemed sufficiently close are merged using the distance function between publications. This involves determining the center of gravity for each cluster and combining clusters whose distance between the centers of gravity does not exceed a specified threshold value.

During the third phase, each cluster is allocated a distinct scientific research focus. To ensure alignment, an expert-driven methodology is employed. This involves making decisions based on the compilation of cluster publications and supplementary details, such as keywords and frequently employed concepts.

In the fourth stage, the authors utilize information about scientists' publication activities, considering the established clusters of scientific directions to which these publications pertain. Solving the problem of identifying the research directions of scientists in the set $A=\{a1,a2,...,an\}$, where n is the number of scientists, involves a mapping function $\Lambda:A\rightarrow V$ where $V=\{\eta_1, \eta_2, \ldots, \eta_\psi\}$ is a set of verbal names representing research directions, and ψ is the number of research directions. The outcomes of this process assist in determining scientists engaged in specific research areas. The discerned areas of scientific research serve as input information for experts forming the project team, taking into account each potential partner's experience in various research directions. The clustering results offer essential insights for addressing partner selection issues, such as employing fuzzy inference.

Identifying potential project groups for implementing research and educational projects based on clustering results.

Let's consider the project in its planning stage after completing the steps of defining the project environment and formulating its objectives. Internal and external factors have been identified, and the project's goals and tasks are clearly outlined. At this juncture, the project is perceived as a series of distinct processes, each dedicated to addressing specific tasks within the overall project framework. The execution of each process necessitates resources, with task performers being a crucial resource among them.

Let's delve deeper into the task of selecting project performers. For simplicity, it is assumed that one performer is assigned to each process from initiation to completion. To streamline the process, the task of selecting project executors into the subtask of choosing precisely one executor for each process is broken down. During the environment determination stage, a set of potential candidates $A=\{a1, a2,...,an\}$ was identified, where n represents the number of potential executors (hereafter referred to as candidates). Each candidate can be evaluated based on criteria $c1, c2, ..., ck$, where k is the number of criteria for evaluating candidates. The objective is to develop a method for evaluating and selecting a suitable performer from the candidate pool, considering various criteria. Furthermore, the method's outcome could be either a single optimal performer a^* or an ordered set of performers $\{a^*1, a^*2, ..., a^*n\}$. The latter approach offers several advantages. In case the most optimal candidate is unable to participate due to unforeseen external factors, the next candidate in line can seamlessly step in.

To assess candidates, one approach is employing multi-criteria group expert evaluation. Utilizing a fuzzy inference system (FIS) is recommended for obtaining generalized aggregated estimates of applicants. The fuzzy inference process involves mapping the vector of input data estimates to an initial scalar value using fuzzy rules.

Fuzzy logical inference involves the following sequential steps:

1. Fuzzification: This initial step relies on linguistic variables and their associated linguistic terms. The primary procedure involves determining the degree of membership of the input value to each linguistic variable.

In the context of the obtained solution, the identification of scientists' research directions can be considered as a discrete fuzzy mapping. The membership function is determined by the ratio

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 16, NOVEMBER 2023

of the number of publications by the author in a specific scientific direction to the total number of publications. Mathematically, it can be expressed as $\Lambda(a_i) = (\eta_b|\mu_b^i), b = \overline{1,\psi}, i = \overline{1,n}$, where b is in the range of 1 to ψ (representing different research directions), and i ranges from 1 to n (representing individual scientists). The membership is determined by the formula (6).

$$\mu_b^i = \frac{\|P(a_i) \cap Y_b\|}{\|P(a_i)\|} \quad (7)$$

where $P(a_i)$ is the set of all publications of the ai scientist, and Yb is a cluster of publications corresponding to the direction of scientific research ηb.

2. Logical Inference Mechanism: This step involves the application of fuzzy rules that specify how input fuzzy sets are mapped to the output fuzzy set. These rules are derived from relevant expert assessments. There are various methods for fuzzy inference, such as the Mamdani, Sugeno, and Larsen approaches.

Fuzzy rules are formulated in the format: "If a candidate possesses competence ηb with a degree of membership μb, then they meet the project requirements with a degree of membership α." These rules help evaluate the suitability of candidates for a specific position based on their competencies. Let's consider an example rule: "If the candidate is knowledgeable in project management methods, then they are exceptionally suitable for the position." In this example, competence in project management methods (ηb) represents the candidate's skills, and "exceptional" is a qualitative assessment. To transform these qualitative assessments into quantitative values represented by membership functions, a specific scale is employed.

The method for translating qualitative expert assessments into membership functions, as described in reference [16], is used to establish this scale. This scale provides a systematic way to transition from qualitative assessments to quantitative degrees of membership. An example of such a scale can be found in Table 1, where different qualitative assessments are mapped to corresponding degrees of membership (e.g., "Poor" corresponds to a degree of membership of 0.1, "Excellent" corresponds to a degree of membership of 0.9). This scale helps standardize the assessment process and ensures that the qualitative evaluations are converted into quantifiable values suitable for fuzzy logic-based evaluations and decision-making.

Table 1 – Scale for verbal assessment of the statement

№	Verbal qualitative assessment	The value of the membership function
1	Great	0,9
2	Well	0,75
3	Satisfactory	0,6
4	Unsatisfactory	0,35

The fuzzy inference procedure involves aggregating all rules. In the Fuzzy Inference System (FIS), the process of fuzzy logical inference consists of determining the degrees of fulfillment for each rule based on the degree of truth of its premise, using the composition $\alpha=\min\{\mu_b\}$. In the Mamdani FIS, the minimum operator is employed, while the Larsen FIS system is based on the product operator.

Defuzzification is the process of converting a fuzzy value into a crisp, clear value. The most commonly used method for defuzzification is the center of gravity method of a fuzzy set.

Defasification occurs using the formula for finding the center of mass

DOI: 10.37943/16AADE3851

$$\frac{\int_{x_{min}}^{x_{max}} x\mu(x)dx}{\int_{x_{min}}^{x_{max}} \mu(x)dx}, \tag{8}$$

where x is a fuzzy quantity, and μ (x) is a function of its membership. Since the authors are dealing with a discrete quantity, the Stiltjes integral should be understood as a sum, that is, defasification occurs according to formula 9.

$$\frac{\sum_{b=1}^{\psi} \alpha^i \mu_b^i}{\sum_{b=1}^{\psi} \mu_b^i} \tag{9}$$

The selection of the project executor involves identifying a scientist whose defuzzification value is maximized. An ordered set of performers is then constructed from scientists in descending order of their defuzzification values.

Conclusions and prospects for further research.

The task of selecting potential partners for cooperation in scientific and educational projects is formulated. A method of selecting project executors is constructed, the essence of which is using fuzzy logical inference to coordinate expert opinions on the requirements for candidates.

The multi-criteria task of selecting potential partners for cooperation in scientific and educational projects has been solved.

In the future, an information technology for forming multi-university scientific and educational communities based on the theory of scientometric analysis will be developed, as well as an information system and an experimental software package that implements it.

Acknowledgment

This paper was written in the framework of the state order to implement the science program according to budget program 217 "Development of Science," IRN No. AP19678627 with the topic: "Development of the information technology for the formation of multi-university scientific and educational communities based on the scientometric analysis theory."

References

[1] Kuchansky, A., Biloshchytskyi, A., Andrashko, Yu., Vatskel, V., Biloshchytska, S., Danchenko, O., & Vatskel, I. (2018). Combined models for forecasting the air pollution level in infocommunication systems for the environment state monitoring. *IEEE 4th Intern. Symp. on Wireless Systems within the Int. Conf. On Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS),* 125–130.

[2] Li, B. & Zhang, J. (2021). A Cooperative Partner Selection Study of Military-Civilian Scientific and Technological Collaborative Innovation Based on Interval-Valued Intuitionistic Fuzzy Set. *Symmetry*, 13, 553. https://doi.org/10.3390/sym13040553

[3] Gladka, M., Kravchenko, O. Hladkyi, Y., & Borashova, S. (2021). Qualification and appointment of staff for project work in implementing IT systems under conditions of uncertainty. *2021 IEEE International Conference on Smart Information Systems and Technologies,* 1-6. https://doi.org/10.1109/SIST50301.2021.9465897

[4] Kolomiiets, A., & Morozov, V. (2021). Investigation of optimization models in decisions making on integration of innovative projects. *Advances in Intelligent Systems and Computing*, 51–64. https://doi.org/10.1007/978-3-030-54215-3_4

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 16, NOVEMBER 2023

[5] Chen, L., Jagota, V. & Kumar, A. (2021). Research on optimization of scientific research performance management based on BP neural network. *Int J Syst Assur Eng Manag*. https://doi.org/10.1007/s13198-021-01263-z
[6] Kuchansky, A., Biloshchytskyi, A., Andrashko, Yu., Biloshchytska, S., Shabala, Ye., & 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), 25–28.
[7] Kuchansky, A., Andrashko, Yu., Biloshchytskyi, A., Danchenko, O., Ilarionov, O., Vatskel, I., & Honcharenko, T. (2018). The method for evaluation of educational environment subjects' performance based on the calculation of volumes of m-simplexes. *Eastern-European Journal of Enterprise Technologies*, 2/4 (92), 15–25.
[8] Biloshchytskyi, A., Kuchansky, A., Paliy, S., Biloshchytska, S., Bronin, S., Andrashko, Yu., Shabala, Ye., & Vatskel, V. (2018). Development of technical component of the methodology for project-vector management of educational environments. *Eastern-European Journal of Enterprise Technologies*, 2/2 (92), 4–13.
[9] Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Yu., & Biloshchytska, S. (2019). Improvement of the method for scientific publications clustering based on n-gram analysis and fuzzy method for selecting research partners. *Eastern-European Journal of Enterprise Technologies*, 4/4 (100), 6–14.
[10] Bykov, V., Biloshchytskyi, A., Kuchansky, A., Andrashko, Yu., Dikhtiarenko, O., & Budnik, S. (2019). Development of information technology for complex evaluation of higher education institutions. *Information Technologies and Learning Tools*, 73(5), 293–306.
[11] Bykov, V., Spirin, O., Biloshchytskyi, A., Kuchansky, A., Dikhtiarenko, O. (2020). Open digital systems for assessment of pegadodical research results. *Information Technologies and Learning Tools*, 75(1), 294–315.
[12] Lizunov P., Biloshchytskyi A., Kuchansky A., Andrashko Yu., Biloshchytska S. The use of probabilistic latent semantic analysis to identify scientific subject spaces and to evaluate the completeness of covering the results of dissertation studies Eastern-European Journal of Enterprise Technologies. 2020. № 4/4 (106). P. 14–20.
[13] Ioannidis, J.P.A., Baas, J. Klavans, R. & Boyack, K. (2019). Supplementary data tables for "A standardized citation metrics author database annotated for scientific field" (PLoS Biology 2019). *Mendeley Data*, 1. https://doi.org/10.17632/btchxktzyw.1
[14] Noorden, R.V. & Chawla, D.S. (2019). Hundreds of extreme self-citing scientists revealed in new database. *Nature*, 572, 578-579. https://doi.org/10.1038/d41586-019-02479-7
[15] Liu, L., & Ran, W. (2020). Research on supply chain partner selection method based on BP neural network. *Neural Comput & Applic*, 32, 1543–1553. https://doi.org/10.1007/s00521-019-04136-6
[16] Newman, M. (2023). Who is the best connected scientist? *A study of scientific coauthorship networks*. https://link.springer.com/chapter/10.1007/978-3-540-44485-5_16
[17] Han, J., Teng, X. & Cai, X. (2019). A Novel Network Optimization Partner Selection Method based on collaborative and knowledge networks. *Information Sciences*, 484. https://doi.org/10.1016/j.ins.2019.01.072
[18] Lungeanu, A., Carter, D., Dechurch, L. & Contractor, N. (2018). How Team Interlock Ecosystems Shape the Assembly of Scientific Teams: A Hypergraph Approach. *Communication Methods and Measures*, 12, 1-25. https://doi.org/10.1080/19312458.2018.1430756
[19] Huilin, X. (2019). Review of methods of evaluation of scientific and research activity for the choice of selection of scientific partners. *Management of development of complex systems*, 38, 156–160. https://doi.org/10.6084/m9.figshare.9788654
[20] Citation Network Dataset: DBLP+Citation, ACM Citation network. (2022). *Aminer*. https://www.aminer.org/citation
[21] Gephi. (2022). *The Open Graph Viz Platform*. https://gephi.org/
[22] I'm knowledge. (2023). *Mendeley*. https://www.mendeley.com/?interaction_required=true