Mathematical Model and Approaches to Quantitative Analysis of Metadata of Scientific Articles*

Hryhorii Hnatiienko[1,*,†], Vitaliy Snytyuk[1,†], Nataliia Tmienova[1,†], Oleksii Ivanchenko[1,†], and Yevhen Patkin[1,†]

[1] *Taras Shevchenko National University of Kyiv, Volodymyrs'ka str. 64/13, Kyiv, 01601, Ukraine*

Abstract

This article is devoted to the study of metadata of scientific articles. Quantitative analysis of this area of research contributes to the structuring of the scientific space. The mathematical model and some approaches to the quantitative analysis of metadata of scientific article are described. The structure of metadata is formalized. In particular, information about authors is structured and the tasks that arise when analyzing this metadata attribute are presented. The structure of the title of a scientific article, its functions, requirements for this attribute, problems of analyzing the titles of scientific articles are considered. The function of membership of the number of words in the title of the article is built. The abstract of a scientific article, its structure, functions of the abstract, requirements for this metadata attribute are investigated, and the problems of analyzing abstracts are presented. The requirements for the set of keywords are also analyzed and the membership function for the number of keywords in a scientific article is constructed. The problems and approaches to the description and analysis of keywords are presented.

Keywords

scientific research, scientific space, scientometric databases, similarity measures, metadata, authors, title, abstract, keywords

1. Introduction

Scientific activity in general, scientific research, and the vast majority of scientific results are unpredictable, and thus the scientific space is a poorly structured area [1, 2]. This phenomenon has been studied from different perspectives for many centuries [3, 4] , but this area of research remains relevant today [5, 6] . The scientific results obtained in the field of research of scientific space [7, 8] contribute to additional structuring of scientific space, its formalization and, thus, to obtaining new knowledge about this important area of cognition.

Scientific publications are one of the indicators of success in the competitive environment of the academic community [9, 10]. The number and quality of scientific publications is an indicator of the scientific level of both each individual scientist and the scientific institution as a whole. Therefore, the study of the relationships associated with scientific publications is a relevant and promising area of analysis [11, 12].

It should be noted that a scientific publication is an electronic or paper publication that promotes the publication of the results of theoretical or experimental research. Such scientific publications are usually intended for professionals and for scientific work. They are the main source of formalized authorship and one of the ways to establish scientific priority.

Information Technology and Implementation (IT&I-2024), November 20-21, 2024, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

g.gna5@ukr.net (H. Hnatiienko); snytyuk@knu.ua (V. Snytyuk); tmyenovox@gmail.com (N. Tmienova); ivanchenko.oleksii@knu.ua (O. Ivanchenko); yevhen.patkin@knu.ua (Y. Patkin)

0000-0002-0465-5018 (H. Hnatiienko); 0000-0002-9954-8767 (V. Snytyuk); 0000-0003-1088-9547 (N. Tmienova); 0000-0002-8526-8211 (Ivanchehko); 0009-0001-2538-1204 (Y. Patkin)

A scientific article is a type of scientific publication that describes a study or a group of studies related to a single topic and is written by scientific authors. A scientific article is one of the most common ways to publish scientific results.

In this article, we will focus primarily on defining the quality and structure of a scientific article. Of course, these aspects are present in any formalized and published scientific article and, surely, they are not the actual scientific result. However, indirectly, scientific results, their recording and popularization largely depend on the correctness of the formal part of scientific publications [13, 14].

2. The main functions of scientific articles

Scientific publications perform several functions [15, 16], which are interrelated and naturally complement each other:

- publishing the results of scientific article;
- recording the completion of a certain stage of research in general;
- confirmation of the fact of approbation or implementation of research results;
- providing primary scientific information to the scientific community;
- notification of the emergence of new scientific knowledge and securing its authorship;
- transfer of new individual scientific results to the public domain;
- serve as a source of data for scientometric databases;
- contribute to raising the formal rating of a researcher;
- provide an opportunity to adequately determine the winners of formal or informal competitions in the educational and scientific environment;
- evidence of the researcher's personal contribution to the development of a scientific problem to determine his relative contribution to the work of the research team;
- ensuring opportunities for transparency and quantitative, reliable reporting;
- satisfy the interests of higher education institutions and academic institutions whose main output is scientific results
- help to establish the author's priority when comparing similar scientific articles, scientific ideas, etc.
- confirm the reliability of the main results and conclusions of the research article, its novelty and scientific level.

3. Metadata structure for scientific articles

Different researchers have some disagreements about which elements of scientific publications should be considered metadata [17, 18]. In this article, we will consider such elements of scientific publications as metadata and comprehensively study them:

- information about the authors (A);
- title of the publication (T);
- abstract (B);
- keywords (W).

The set of metadata of scientific articles can be represented as a tuple:

$$\langle A,T,B,W \rangle. \quad (1)$$

The metadata of a scientific article is an important part of scientific research: they are symbiotic in nature, and the importance of each of the above elements is unique in its own way. The absence of any of the above metadata elements has a significant impact on the quality of a scientific publication, and incorrect metadata design significantly affects the effectiveness of the entire scientific publication.

In general, metadata provides an opportunity for additional analysis of the scientific space and helps to structure research papers, turning a poorly structured subject area into a structured one [1, 19].

3.1. Problems of metadata analysis

In many practical cases, there is a need to analyze all metadata components. And with regard to the first three components of tuple (1), algorithms for comparing, determining preference relations, their metricizing, calculating similarity measures of the components of tuple (1), etc. are often used [20].

These are just some of the common problems that are solved with the help of metadata or on the basis of metadata-related information:

- making a decision on the official indexing of a scientific paper in a scientometric database [21];
- facilitating automated marketing of authors of well-described and properly formatted research papers;
- solving the problem of reviewing articles in scientific journals or appointing reviewers to select materials for international conferences, symposia, etc.;
- justified selection of scientists to evaluate projects submitted to research project competitions, student research competitions, etc.;
- determination of opponents and reviewers of one-time specialized academic councils when awarding the measure of Doctor of Philosophy [23, 24]
- in many other important activities and events that require formalization and justification of the choice [25, 26].

4. Structure of information about authors

Information about the authors (A) is a metadata attribute and usually contains the following basic characteristics $a_i \in A,\ a_i = (a_i^1,...,a_i^6),\ i \in I$, where

$I -$ a set of indexes of information about authors;

a_i^1 – the last name and first name or initials of the author or authors;

a_i^2 – academic measure, academic title, honorary title, etc.;

a_i^3 – data on the authors' affiliations: position and official name of the institution (enterprise, organization) at the main place of work, city, country, e-mail address, home and mobile phone numbers of the key author, etc.;

a_i^4 – ORCID identifiers;

a_i^5 – author profiles in scientometric databases;

a_i^6 – other information about the authors.

4.1. Problems of analyzing information about authors

When analyzing the characteristics of authors, a number of problems arise that should be formalized and solved in order to structure the scientific space and solve other actual problems. Such

problems are solved, in particular, to improve the scientific space and develop open science trends [27, 28]. In particular, the following problems can be solved:

- clustering of authors of scientific articles based on the analysis of their profiles in scientometric databases;
- identification of informal research groups based on the analysis of author profiles and other indicators;
- determining the level and number of indicators of mutual influence in informal research groups;
- determining the dynamics of authors' publications, their productivity, publication activity, creativity, etc.;
- detecting cases of popularity manipulation based on the analysis of author profiles.

5. Structure of the title of scientific article

The title of a scientific article (T) is a metadata attribute and, in turn, has a number of characteristics that will be discussed below.

5.1. Problems of metadata analysis

It is known that in 70% of cases when readers first read a scientific paper, its title is the only factor that focuses the attention of readers of scientific articles. It is believed that the probability that a user will skip to the next paragraph of the article after the title if it does not attract his attention is reduced by 10 times. So, the following heuristic can be derived from this.

Heuristic H1. For every person who reads the entire article, there are hundreds of people who read the title without looking at the article.

The title is also the main search criterion in scientometric databases, i.e., we have such ranking by a set of metadata attributes (1) in situations of determining the impact of metadata in studies with an undirected search for information sources:

$$T \succ W \succ B \succ A. \tag{2}$$

We will assume that $\tau_i \in T, i \in I = \{1,...,n\}$, where $n-$ is the number of studied titles.

All metadata attributes of a scientific article are important, but from the point of view of qualitative preparation of the article for publication, from the point of view of its correct formatting, from the point of view of its indexing in scientometric databases and the reliability of using the scientific potential of the article for its dissemination in the scientific space, the most adequate is the ranking of attributes of type (2) [29, 30]. After all, most bibliographic search engines, databases, or journal websites rely heavily on the title of the article when implementing the search algorithm.

Here are the criteria for a good title name:

$f_1(\tau_i)$ – relevance of the article's concept to the subject matter of the publication;

$f_2(\tau_i)$ – title reflects the subject of the study, not just the result;

$f_3(\tau_i)$ – no "screaming" headlines in the style of the "yellow press" are used;

$f_4(\tau_i)$ – no verbs are used;

$f_5(\tau_i)$ – no vulgarities are used;

$f_6(\tau_i)$ – the content of the article is described as accurately and discreetly as possible;

$f_7(\tau_i)$ – the essence of the paper is concisely stated;

$f_8(\tau_i)$ – matching the length of the title to the editorial requirements.

It is believed that a successful title of a scientific article guarantees interest in this article on the part of readers and the corresponding intensity of dissemination of the article in the scientific community.

5.2. Requirements for the title of a scientific article

The title of a scientific article should clearly define the main purpose of the article, begin with the name of the research object, and attract the attention of a potential reader. A properly formed title of any scientific article $\tau_i \in T, i \in I,$ should meet at least several reasonable requirements. When formulating a title, you should be guided by certain principles, including the following:

$g_1(\tau_i)$ – adequacy of the article's content;

$g_2(\tau_i)$ – specificity of the wording;

$g_3(\tau_i)$ – informativeness of the wording (the measure of informativeness will be developed by the authors in further research);

$g_4(\tau_i)$ – absence of such phrases as: "Research of the question", "Research problems", "Some research questions", "Ways of solving", etc.;

$g_5(\tau_i)$ – no abbreviations and acronyms other than commonly used ones;

$g_6(\tau_i)$ – clarity, capacity and conciseness of the main idea of the article in the title;

$g_7(\tau_i)$ – absence or minimization of keywords in the title of the article in order to increase the number of indicators for indexing the article in scientometric databases;

$g_8(\tau_i)$ – correspondence between the title and purpose of the article, the main results and conclusions (the measure of such correspondence will also be developed by the authors in further research);

$g_9(\tau_i)$ – absence of common phrases in the title name;

$g_{10}(\tau_i)$ – observance of the correct word order: important words should always be listed in the title in the first place.

Following the principles of $g_1(\tau_i) - g_{10}(\tau_i), i \in I,$ the researchers believe that a good title of a scientific article increases the number of readers of a scientific article, its scientometric indicators, increases the popularity of the authors of a scientific article in the scientific space, etc.

A good title for a scientific article should contain the following components:

τ_i^R – outcome (in particular, principles, models, methods, classifications, information technology, constructed clustering, etc.);

τ_i^O – the object of research;

τ_i^N – brief information about the scientific novelty, i.e., what distinguishes it from all other scientific articles and highlights what has not been done before.

That is, a good title of the article should be expressed by a formula:

$$\tau_i = \tau_i^R + \tau_i^O + \tau_i^N, i \in I, \tag{3}$$

Moreover, the "+" sign in formula (3) reflects the concatenation of the components of the title of a scientific article.

5.3. The problem of analyzing the titles of scientific articles

To formalize the research of the title of a scientific article, we can formulate and study different types of problems that are formalized in different classes of mathematical models. In particular, the following aspects of the title of a scientific article should be investigated:

- name structure;
- the title corresponds to the content of the scientific article;
- matching the length of the article title to a high value of the membership function that will be built below;
- the article and its title belong to the research areas declared by the authors according to the classification of scientometric databases or other indicators;
- the quality of the article's design in terms of external attributes, in particular, the requirements for the title.

In order to obtain high-quality results in solving these problems, it is necessary to choose the appropriate areas of research.

5.4. Length of the title of the scientific article

Some values are inappropriate and incorrect to represent in a fixed form. By their construction and function, they are interval. Limiting the intervals in such cases is violence against experts and leads to deliberate a priori inaccuracy in expert assessments, i.e. deliberate conformism of experts.

To determine the recommended number of words in the title of a scientific article, the authors of this paper conducted a computational experiment. It should be emphasized that the sample is not representative, but it largely indicates trends in the requirements of scientific editors of popular and important scientific journals for recommendations on the length of article titles.

The authors of this article have studied 55 websites that contain information on the rules for formatting scientific articles. Based on the analysis of sites found by the Google search engine using the queries "Number of words in the title of a scientific article", "Length of the title of a scientific article", "Requirements for the title of a scientific article", Table 1 was compiled.

Table 1
The recommended number of words in the title of a scientific article with incomplete information

Number in order	Lower recommended interval boundary	Upper recommended interval boundary	Number of sites with such recommendations
1	5	10	3
2	3	7	7
3	3	10	1
4	5	6	2
5	5	8	3
6	5	9	1
7	5	7	5
8	uncertainty	8	3
9	uncertainty	12	1
10	uncertainty	13	1
11	uncertainty	15	2
12	5	uncertainty	15
13	uncertainty	4	6
14	7	15	2
15	uncertainty	5	2
16	10	15	1

It is obvious that the presentation of information in Table 1 is not only incomplete, but also contains some "exotic" requirements for the number of words in the title. We will not exclude any of the editors' suggestions from the generated data set and will introduce appropriate heuristics for greater certainty.

Heuristic H2. To fill in row 8 of Table 1, we assume that the lower bound of the recommended word count is 5. For rows 9 and 10, the lower bound is 6, and for row 11, the lower bound is 7. For lines 13 and 15, the lower bound is 3.

Heuristic H3. For row 12 of Table 1, the upper bound is 13.

Supplementing Table 1 with heuristic H2 and heuristic H3 and arranging the boundaries of the intervals presented in the table in ascending order, we obtain Table 2.

Table 2

The recommended number of words in the title of a scientific article with supplemented and organized information

Number in order	Lower recommended interval boundary	Upper recommended interval boundary	Number of sites with such recommendations
1	3	4	6
2	3	5	2
3	3	7	7
4	3	10	1
5	5	6	2
6	5	7	5
7	5	8	6
8	5	9	1
9	5	10	3
10	5	13	15
11	6	12	1
12	6	13	1
13	7	15	4
14	10	15	1

Note 1: The number of rows in Table 2 has decreased compared to the number of rows in Table 1 because after the introduction of the H2 heuristic and the H3 heuristic, the values of some rows coincided and the corresponding rows were merged.

To visualize the information in Table 2, let's plot the number of words in the title recommended by editors and present it in Figure 1. The numbers in the cells on the orange background reflect the frequency of use of the corresponding interval on websites based on queries to the Google search engine.

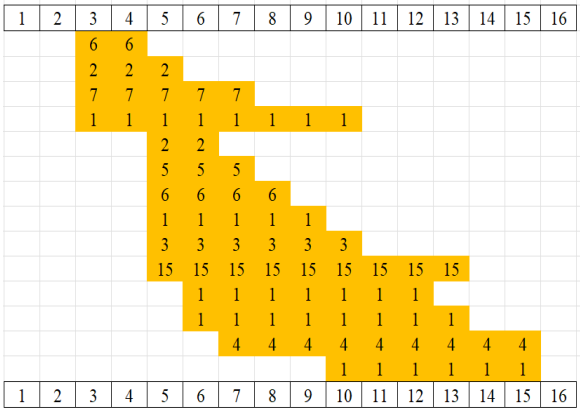

Figure 1: Illustration of the information on the recommended number of words in the title of a scientific article presented in Table 2

To build a membership function for the number of words in the title of a scientific article, we applied the layering method using the data presented in Table 2. The results of the layering method are approximated by a trapezoidal membership function. The results of this procedure are shown in Figure 2. In this case, the maximum values of the membership function are in the interval with the boundaries of 6-7 words in the title of the scientific article.

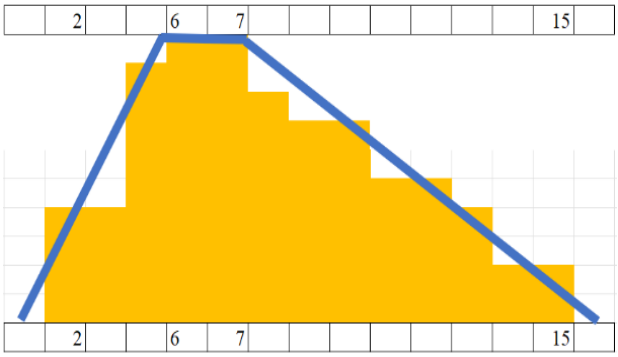

Figure 2: Results of building a membership function for the number of words in the title of a scientific article

The membership function of the indicator "Number of words in the title of a scientific article", calculated as a trapezoid, is as follows:

$$\mu_T(\tau,1,6,7,15)=\begin{Bmatrix} 0,\ \tau\le 1, \\ (\tau-1)/5,\ 1\le\tau\le 6, \\ 1,\ 6\le\tau\le 7 \\ (15-\tau)/8,\ 7\le\tau\le 15, \\ 0,\ \tau\ge 15 \end{Bmatrix}.$$

If the membership function is constructed based on the frequency of values, as shown in Figure 1, i.e., based on the number of sites represented in the last column of Table 2, the membership function will look like Figure 3. In this case, the best approximation is the triangular membership function. The membership function reaches its maximum value at point 7, which corresponds to the number of words in the title of the scientific article.

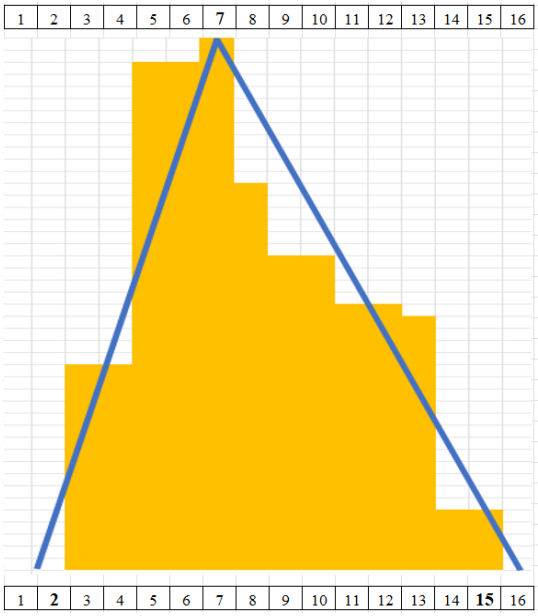

Figure 3: The results of building a membership function for the number of words in the title, taking into account the frequency of values

The formula for the triangular membership function for the number of words in the title of a scientific article shown in Figure 3 is as follows:

$$\mu_{\Delta}(\tau,1,7,15)=\begin{Bmatrix} 0,\ \tau\le 1, \\ (\tau-1)/6,\ 1\le\tau\le 7, \\ (15-\tau)/8,\ 7\le\tau\le 15, \\ 0,\ \tau\ge 15 \end{Bmatrix}.$$

6. Research of Abstract of the scientific article

The Abstract is an important attribute of a scientific article and plays an important role in the realization of the main functions of a scientific publication described in Section 2 of this article. According to the state of research on this metadata attribute, the Abstract (B) is more important than the Author Information (A) and inferior to the Title (T) and Keywords (W), as reflected in the ranking (2). This situation indicates, first of all, the need for further in-depth study of this metadata attribute of a scientific article.

6.1. Structure, functions, and requirements for abstract

An important indicator of an abstract of a scientific article is its structure. In general, the structure of an abstract should include:

- information about the main ideas and conclusions of the research;
- a brief description of the main content of the article;
- a list of the main issues covered in the article;
- information about the purpose of the article and its scientific value;
- subject of the research;
- formulation of the problem;
- ways to solve the problem;
- the topic of a scientific article;
- the purpose of the article;
- description of methods, research methodology or methodological justification;
- information indicating the relevance of the issue;
- scientific novelty of the work and a description of its scientific significance;
- a brief description of the experimental studies;
- main results of the research;
- the scope of the results;
- description of the practical significance of the results and their practical value;
- brief conclusions about the research work.

The structure of an abstract may contain a different number of the above components in a different order. However, the success of a research paper largely depends on the correct structure of the abstract. An abstract includes the following attributes:

b_j^1 – presents a description of the main topic;

b_j^2 – reflects the problem discussed in the article;

b_j^3 – focuses on the research object presented in the article;

b_j^4 – summarizes the purpose of the research work;

b_j^5 – presents and announces the main results of the work;

b_j^6 – contains the scientific novelty of the study in comparison with other scientific works related to the topic and purpose of the research;

b_j^7 – gives the reader a complete picture of the content of the research paper

b_j^8 – conveys information about the article, not just a summary of that information.

A well-prepared abstract should contain all these characteristics: $b_j = \left(b_j^1,...,b_j^8\right), j \in J.$

The abstract must additionally satisfy the following heuristic.

Heuristic H4. An important characteristic of a well-designed scientific article is a certain predefined level of similarity between the abstract (B) and the article content, title (T), and keywords (W).

At the same time, an important requirement for this metadata attribute is that the content of the abstract should not coincide with the text of the main part of the article or the conclusions; the purpose of the abstract is to attract the reader's attention.

Heuristic H5. The abstract is read by at least 100 times more people than the scientific article itself.

The requirements of different editorial boards for abstracts are varied and have significant differences. The length of the abstract depends on the requirements of the editorial board, the field of science, the direction of research, etc. In this case, it is impractical to build a membership function for the length of the abstract. However, the information on the size of abstracts studied by the authors of this paper is summarized in Table 3.

Table 3

Recommended number of words in the abstract of a scientific article

Number in order	Lower recommended interval boundary	Upper recommended interval boundary	Number of sites with such recommendations
1	uncertainty	50	
2	100	150	Exact sciences
3	150	250	
4	150	300	Humanities
5	200	250	
6	uncertainty	500 characters	1

Note 2: Some editorial boards, for example, for articles in the humanities, set a maximum length limit for abstracts and keywords of 1800 characters. Other editorial boards set a range of 1800-2500 characters. Other quantitative values are also found among the requirements.

6.2. Problems of abstract analysis

As noted above, the abstract (B) as an attribute of the metadata of a scientific article is currently the least formalized and its analysis can be largely automated in the future. For this purpose, the following problems should be formulated and solved, in particular:

- defining tools for creating an abstract;
- studying the similarity measure between the existing abstract of a published article and the automatically generated one;
- generating text abstracts using different approaches and determining the similarity measure between the generated abstracts in order to identify the tools that best generate text abstracts in the selected field of knowledge;
- using abstracts to find reviewers for a journal or conference;

- determining whether the abstract reflects the content of the article in quantitative terms
- calculating the quantitative correlation between the abstract and the title of the article;
- calculation of quantitative correlation indicators between the abstract and the keywords of the article;
- formalizing the condition and determining whether the abstract meets the restrictions of the editorial board;
- determining the direction of the research and deciding the area in which the article need to be included, in particular, building the appropriate membership function;
- monitoring the existing similarity measures between texts based on abstracts and developing new similarity measures, if necessary;
- classification by abstracts of the materials submitted to the scientific event [31] of publications in accordance with the declared areas (sections) of the scientific event.

7. Researching a set of keywords for a scientific article

To further study this metadata attribute of a scientific article, it is advisable to first introduce a definition of this attribute.

Definition. A keyword is a word or a stable phrase from the text of an article that carries a semantic load in terms of information retrieval. The set of keywords should reflect the main content of the scientific article out of context. In addition, keywords should be specific, relevant to the subject area under study, meaningful, and unique.

Keywords are intended for the efficient use of search engines and the systematization of articles by topics. Keywords are used to index scientific articles in databases that abstract the material.

Keywords can also be used to build a citation index. Let's formulate the basic requirements for keywords:

w_i^1 – avoid general and ambiguous terms;

w_i^2 – avoid using abbreviations that are commonly used in the relevant field of research;

w_i^3 – keywords and words from the title should not repeat each other, since both of these elements are simultaneously specified in databases.

7.1. Research the recommended number of keywords

Based on the analysis of 25 websites found by the Google search engine using the queries "Number of keywords for a scientific article", "Keywords for a scientific article", "Requirements for the number of keywords for a scientific article", Table 4 was compiled.

Note 1: There are cases when editorial policy excludes keywords altogether.

Remark 2: The ambiguity is, in particular, due to the fact that in some publications the phrase is considered to be one keyword.

For the sake of certainty, we will supplement the upper or lower bounds of intervals that are not determined by the editorial boards by introducing appropriate heuristics.

Heuristic H6. To fill in rows 7 and 10 of Table 4, we assume that the lower bound of the recommended number of keywords is 2. For row 17, the lower bound is 4, and for row 22, the lower bound is 6.

Heuristic H7. For row 1 of Table 4, we set the upper bound to 6. For row 11, the upper bound is 9, and for row 23, the upper bound is 13.

Supplementing Table 4 with heuristic H6 and heuristic H7 and arranging the boundaries of the intervals presented in the table in ascending order, we obtain Table 5.

Note 3. The number of rows in Table 5 has decreased compared to the number of rows in Table 4 after the introduction of the H6 heuristic and the merging of rows with the same limit values.

Based on Table 5, we use the layering method to build a membership function for the recommended number of keywords in a scientific article. The constructed geometric figure is approximated by a trapezoidal membership function, as shown in Figure 3.

Table 4
Recommended number of keywords in a scientific article

Number in order	Lower recommended interval boundary	Upper recommended interval boundary	Number of sites with such recommendations
1	3	uncertainty	1
2	3	4	1
3	3	5	1
4	3	7	3
5	3	8	1
6	3	10	6
7	uncertainty	4	1
8	4	6	2
9	4	8	3
10	uncertainty	5	1
11	5	uncertainty	17
12	5	6	1
13	5	7	5
14	5	8	5
15	5	10	7
16	5	15	19
17	uncertainty	6	5
18	6	8	2
19	6	9	1
20	7	10	1
21	8	12	1
22	uncertainty	10	4
23	10	uncertainty	1
24	10	12	1
25	10	15	1

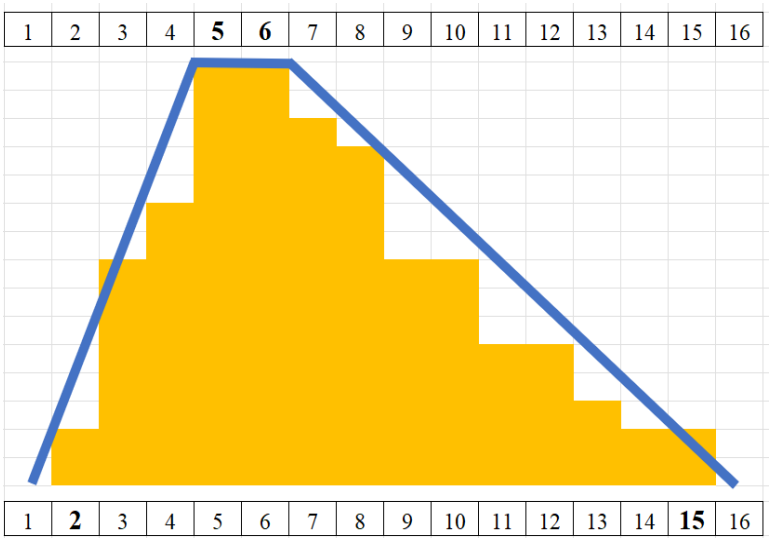

Figure 4: Membership function for the number of keywords in a scientific article without taking into account the frequency of values

The formula for the trapezoidal membership function shown in Figure 4 is as follows:

$$\mu_T(w,1,5,6,15)=\begin{Bmatrix} 0,\ w\le 1, \\ (w-1)/4,\ 1\le w\le 5, \\ 1,\ 5\le w\le 6 \\ (15-w)/9,\ 6\le w\le 15, \\ 0,\ w\ge 15 \end{Bmatrix}.$$

Let's also build a membership function taking into account the frequency of values, i.e., taking into account the number of sites presented in the last column of Table 5. In this case, the membership function will look like the one shown in Figure 4. The best approximation for the figure is a triangular membership function. This membership function reaches its maximum value at point 6, which corresponds to the number of keywords in the scientific article.

Table 5
Recommended number of keywords in a scientific article

Number in order	Lower recommended interval boundary	Upper recommended interval boundary	Number of sites with such recommendations
1	2	4	1
2	2	5	1
3	3	4	1
4	3	5	1
5	3	6	1
6	3	7	3
7	3	8	1
8	3	10	6
9	4	6	7
10	4	8	3
11	5	9	17
12	5	6	1
13	5	7	5
14	5	8	5
15	5	10	7
16	5	15	19
17	6	8	2
18	6	9	1
19	6	10	4
20	7	10	1
21	8	12	1
22	10	12	1
23	10	13	1
24	10	15	1

The formula for the triangular membership function shown in Figure 5 is as follows:

$$\mu_\Delta(w,1,6,15)=\begin{Bmatrix} 0,\ w\le 1, \\ (w-1)/5,\ 1\le w\le 6, \\ (15-w)/9,\ 6\le w\le 15, \\ 0,\ w\ge 15 \end{Bmatrix}.$$

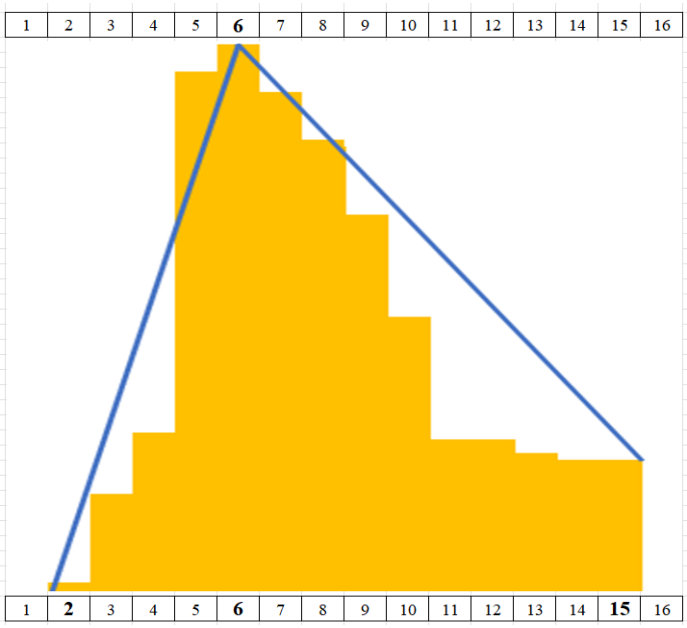

Figure 5: Membership function for the number of keywords in an article based on the frequency of values

7.2. Problems and approaches to describing and analyzing keywords

For the formalized description and analysis of keywords of scientific articles as an attribute of tuple (1), we will introduce additional heuristics.

Heuristic H8. The keywords sufficiently reflect the content and focus of the scientific article and can be used to determine the similarity of research areas in other scientific papers.

Heuristic H9. The similarity measure between the sets of keywords of any two scientific articles sufficiently reflects the similarity of the content of these articles (for some areas of research and decision-making situations).

Heuristic H10. The study of the similarity of keyword sets can be used to identify clusters of research groups and to identify the similarity of research interests of researchers.

Keywords are used by search engines to index and rank scientific articles in search results and it is an effective tool for increasing citations.

An important characteristic of a set of keywords defined by the author or with the help of software tools is its relevance, i.e. the level of correspondence between the specified set and the full text of the scientific article that this set of words represents.

The main criteria for determining the quality of keyword research are as follows:

$\varphi_1(w_i)$ – accuracy in defining the topic of the research article and the essence of the article, conference materials, etc.;

$\varphi_2(w_i)$ – quality of the characterization of the field of scientific research;

$\varphi_3(w_i)$ – the potential to help to group the information that researchers are looking for;

$\varphi_4(w_i)$ – the ability to guide authors in finding the scientific material they need;

$\varphi_5(w_i)$ – speeding up the process of searching and classifying information about the subject of research;

$\varphi_6(w_i)$ – giving a potential reader an idea of the article;

$\varphi_7(w_i)$ – the number of keywords in the article.

Heuristic H11. According to statistics, only every tenth user who reads the keywords then goes to the text of the article.

You should choose words that are frequently repeated in the main material, with the exception of stop words, i.e. words that do not carry a semantic load. Stop words include particles, prepositions, interjections, pronouns, introductions, etc.

At the same time, next heuristic is fair.

Heuristic H12. The importance and decision-making process for including words or phrases in a set of keywords for a scientific article can be determined by other criteria than just frequency.

8. Conclusions.

This article discusses and studies various aspects of metadata of scientific articles. The problem of formalizing metadata as an integral part of the scientific space is investigated. Thus, this area of research is transformed from a poorly structured one into a structured one. In particular, the elements of metadata are studied: information about authors, article title, abstract, keywords. Membership functions were also constructed based on computational experiments conducted by the authors.

The research area proposed by the authors of this article has broad prospects for development [32, 33]. Further work can explore the possibilities of manipulating metadata, approaches to the use of natural language processing technologies to automate the processes of improving the quality of metadata, and can develop algorithms for identifying and correcting errors in the design of a scientific article and its structural elements, methods for automatically improving the quality of scientific article design, determining the quality of metadata preparation in terms of its indexing, etc. [34].

Neural networks can also be used to identify potential situations of academic dishonesty, cases of manipulation of authors' popularity, etc. Such results can be used to create expert groups in scientific fields, identify reviewers for student research competitions, competitions of projects of the Ministry of Education and Science, various tenders, competitions of infrastructure development projects, identify the best urban transformation projects, form specialized scientific councils for the defense of theses, appoint reviewers for articles in journals, conferences, etc.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Rabin, A. V., Petrushevskaya, A. A., & Sinitsin, O. V. (2020). Methods and formal models of intelligent analysis of weakly structured data. IOP Conference Series: Materials Science and Engineering, 734(1), 012159. https://doi.org/10.1088/1757-899X/734/1/ 012159

[2] Eberendu, A. (2016). Unstructured Data: An overview of the data of Big Data. International Journal of Computer Trends and Technology, 38, 46–50. https://doi.org/10.14445/22312803/IJCTT-V38P109

[3] Jonassen, D.H. Instructional design models for well-structured and III-structured problem-solving learning outcomes. ETR&D 45, 65–94 (1997). https://doi.org/10.1007/BF02299613

[4] Borozdykh N.V. Principles of open science as the basis for the formation of the scientific space in Ukraine. Science and scientific research. 2023. No. 2 (120). P. 116–137. https://doi.org/10.15407/sofs2023.02.116 {in Ukrainian]

[5] Uriarte, Filemon A. (2008). Knowledge Management: Konsep, Arsitektur dan Impelementasi Introduction to Knowledge Management: Abrief introduction to the basic elements of knowledge management for non-practitioners interested in understanding the subject. Jakarta: ASEAN Foundation.

[6] Edopkolor, J. E., & Osifo, K. E. (2022). Knowledge management and job performance of business studies teachers: The mediating effect of work engagement. Management Review: An International Journal, 17(1), 27-64.
[7] Hirsch J.E. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences. 2005; 102:16569–16572 https://doi.org/10.1073/pnas.0507655102 PMID: 16275915
[8] Ruan X, Ao W, Lyu D, Cheng Y, Li J. Effect of the topic-combination novelty on the disruption and impact of scientific articles: Evidence from PubMed. Journal of Information Science. 2023; https://doi. org/10.1177/01655515231161133
[9] The Leiden Manifesto for research metrics. Nature. 2015. Vol. 520. PP. 429- 431. URL: https://www.researchgate.net/publication/275335177. DOI: 10.1038/520429a.
[10] Mukherjee D, Lim WM, Kumar S, Donthu N. Guidelines for advancing theory and practice through bibliometric research. J Bus Res. (2022) 148:101–15. doi: 10.1016/j.jbusres.2022.04.042
[11] Li M, Livan G, Righi S (2024) Breaking down the relationship between disruption scores and citation counts. PLoS ONE 19(12): e0313268. https://doi.org/10.1371/journal.pone.0313268
[12] Veugelers R, Wang J. Scientific novelty and technological impact. Research Policy. 2019; 48:1362– 1372 https://doi.org/10.1016/j.respol.2019.01.019
[13] Radicchi F, Fortunato S, Castellano C. Universality of citation distributions: Toward an objective measure of scientific impact. PNAS. 2008; 105:17268–17272 https://doi.org/10.1073/pnas.0806977105 PMID: 18978030
[14] Bu Y, Waltman L, Huang Y. A multi-dimensional framework for characterizing the citation impact of scientific publications. Quantitative Science Studies. 2021; 2:155–183 https://doi.org/10.1162/qss_a_ 00109
[15] Abramo G, D'Angelo CA, Rosati F. Career advancement and scientific performance in universities. Scientometrics. 2014; 98:891–907 https://doi.org/10.1007/s11192-013-1075-8
[16] Hnatiienko, H., Snytyuk, V., Tmienova, N., Voloshyn, O. Determining the effectiveness of scientific research of universities staff / CEUR Workshop Proceedings, Volume 2833, 2021, Pages 164-176 // 7th International Conference ""Information Technology and Interactions"", IT and I 2020; Kyiv; Ukraine; 2 December 2020 through 3 December 2020; Code 167962
[17] Virkus, S., Garoufallou, E.: Data Science from a Perspective of Computer Science. In: Garoufallou, E., Fallucchi, F., William De Luca, E. (eds.) Metadata and Semantic Research. pp. 209–219. Communications in Computer and Information Science, Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-36599-8_19
[18] J. Greenberg, H.C. White, S. Carrier and R. Scherle, A metadata best practice for a scientific data repository, Journal of Library Metadata 9(3) (2009), 194–212.
[19] Hryhorii Hnatiienko, Georgii Gaina, Oleh Ilarionov, Vitaliy Snytyuk and Nataliia Tmienova. Methods of Identifying the Correlation of Ukrainian Scientific Paradigms Based on the Study of Defended Dissertations / CEUR Workshop Proceedings, Volume 3646, Pages 64 – 75, 2023 // Selected Papers of the X International Scientific Conference "Information Technology and Implementation" (IT&I 2023). Workshop Proceedings Kyiv, Ukraine, November 20-21, 2023.
[20] Voloshin, A.F., Gnatienko, G.N., Drobot, E.V. A Method of Indirect Determination of Intervals of Weight Coefficients of Parameters for Metricized Relations Between Objects // Journal of Automation and Information Sciences, 2003, 35(1-4). DOI: 10.1615/JAutomatInfScien.v35.i3.30
[21] Hryhorii Hnatiienko, Dmytro Nelipa, Vitaliy Snytyuk, Nataliia Tmienova, Oleksii Voloshyn. The Method of Determining the Index of Geographical Representation of a Scientific Event / CEUR Workshop Proceedings, Volume 3538, Pages 181-197, 2022 // Selected Papers of the III International Scientific Symposium "Intelligent Solutions" (IntSol-2023). Symposium Proceedings Kyiv - Uzhhorod, Ukraine, September 27-28, 2023.
[22] Hnatiienko, H., Snytyuk, V., Tmienova, N., Voloshyn, O. Application of expert decision-making technologies for fair evaluation in testing problems // Selected Papers of the XX International

Scientific and Practical Conference "Information Technologies and Security" (ITS 2020), Kyiv, Ukraine, December 10, 2020 / CEUR Workshop Proceedings, 2021, 2859, pp. 46–60.
[23] Meltem Aksoy, Seda Yanik, and Mehmet Fatih, Amasyali. Reviewer Assignment Problem: A Systematic Review of the Literature. J. Artif. Int. Res. 76, May 2023. https://doi.org/10.1613/jair.1.14318 .
[24] Petrichko M. V., Shtovba S. D. (2024). Automation of the selection of scientific reviewers: a review of tasks and methods, Bulletin of Vinnytsia Polytechnic Institute, No. 1, pp. 56–64. https://doi.org/10.31649/1997-9266-2024-172-1-56-64
[25] Serhiy Shtovba, Mykola Petrichko. Express selection of opponents for one-time defense councils of PhD dissertations // Ukrainian Journal of Information Systems and Data Science Volume 2, Issue 1, 2024.
[26] Hryhorii Hnatiienko, Oleksii Hnatiienko, Oleh Ilarionov, Oleksii Ivanchenko, Vitaliy Snytyuk. The Method of Determining the Priority of Candidates by Means of Preferential Voting Based on an Algebraic Approach / CEUR Workshop Proceedings, Volume 3538, Pages 232-244, 2023 // Selected Papers of the III International Scientific Symposium “Intelligent Solutions” (IntSol-2023). Symposium Proceedings Kyiv - Uzhhorod, Ukraine, September 27-28, 2023.
[27] Nosenko, Y., & Sukhikh, A. (2020). Open science in the context of building a knowledge society and digital transformation of the European space. Physical and Mathematical Education, 4 (26), 85—92 [in Ukrainian].
[28] Lokshyna, O. (2018). Open education in the European space: a strategy for development. Pedagogical Sciences: Theory, History, Innovative Technologies, 2, 75—86 [in Ukrainian].
[29] Tsyganok V.V., Kadenko S.V., Andriichuk O.V. Considering Importance of Information Sources during Aggregation of Alternative Rankings / CEUR Workshop Proceedings, Vol. 2067. Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017); Kyiv, Ukraine, November 30, 2017. P.132-141.
[30] Hnatiienko H., Tmienova N., Kruglov A. (2021) Methods for Determining the Group Ranking of Alternatives for Incomplete Expert Rankings. In: Shkarlet S., Morozov A., Palagin A. (eds) Mathematical Modeling and Simulation of Systems (MODS'2020). MODS 2020. Advances in Intelligent Systems and Computing, vol 1265. Springer, Cham. https://doi.org/10.1007/978-3-030-58124-4_21. Pp. 217-226.
[31] Hnatiienko, H., Snytyuk, V., Tmienova, N. Calculation of the integral quality index of a scientific event in the context of the interests of a scientific institution // Selected Papers of the XXI International Scientific and Practical Conference "Information Technologies and Security" (ITS 2021), Kyiv, Ukraine, December 9, 2021 / CEUR Workshop Proceedings, 2021, 3241, pp. 79–91.
[32] A. Nisioti, G. Loukas, A. Laszka, E. Panaousis Data-driven decision support for optimizing cyber forensic investigations. IEEE Trans. Inf. Forensics Secur., 16 (2021), pp. 2397-2412, https://doi.org/10.1109/TIFS.2021.3054966
[33] Michael C. Todd, Gilbert L. Peterson, "Temporal metadata analysis: A learning classifier system approach", Forensic Science International: Digital Investigation, vol.51, pp.301842, 2024.
[34] Rodrigues, J., & Teixeira Lopes, C. (2023). Images as Metadata: A New Perspective for Describing Research Data. Journal of Library Metadata, 23(3–4), 87–101. https://doi.org/10.1080/19386389.2023.2252722