The methods for identification of near-duplicates in electronic scientific papers, which include the content of the same type, for example, text data, mathematical formulas, numerical data, etc. were described. For text data, the method of locally sensitive hashing with the finding of Hamming distance between the elements of indices of electronic scientific papers was formalized. If Hamming distance exceeds a fixed numerical threshold, a scientific paper contains a near-duplicate. For numerical data, sub-sequences for each scientific work are formed and the proximity between the papers is determined as the Euclidian distance between the vectors consisting of the numbers of these sub-sequences. To compare mathematical formulas, the method for comparing the sample of formulas is used and the names of variables are compared. To identify near-duplicates in graphic information, two directions are separated: finding key points in the image and applying locally sensitive hashing for individual pixels of the image. Since scientific papers often include such objects as schemes and diagrams, subscriptions to them are examined separately using the methods for comparing text information. The combined method for identification of near-duplicates in electronic scientific papers, which combines the methods for identification of near-duplicates of various types of data, was proposed. To implement the combined method for the identification of near-duplicates in electronic scientific papers, an information-analytical system that processes scientific materials depending on the content type was devised. This makes it possible to qualitatively identify near-duplicates and as widely as possible identify possible abuses and plagiarism in electronic scientific papers: scientific articles, dissertations, monographs, conference materials, etc.

Keywords: near-duplicate, electronic scientific paper, antiplagiarism system, locally sensitive hashing

UDC 005.8
DOI: 10.15587/1729-4061.2021.238318

DEVELOPMENT OF THE COMBINED METHOD OF IDENTIFICATION OF NEAR DUPLICATES IN ELECTRONIC SCIENTIFIC WORKS

Petro Lizunov
Doctor of Technical Sciences, Professor, Head of Department
Department of Fundamentals of Informatics
Kyiv National University of Construction and Architecture
Povitroflotskyi ave., 31, Kyiv, Ukraine, 03037
Andrii Biloshchytskyi
Doctor of Technical Sciences, Professor
Astana IT University
Mangilik Yel ave., EXPO Business Center, Block C.1.,
Nur-Sultan, Republic of Kazakhstan, 010000
Department of Information Systems and Technologies*
Alexander Kuchansky
Corresponding author
Doctor of Technical Sciences, Associate Professor
Department of Information Systems and Technologies*
E-mail: kuczanski@gmail.com
Yurii Andrashko
PhD, Associate Pofessor
Department of System Analysis and Optimization Theory
Uzhhorod National University
Narodna sq., 3, Uzhhorod, Ukraine, 88000
Svitlana Biloshchytska
Doctor of Technical Sciences, Associate Professor
Department of Intelligent and Information Systems*
Oleg Serbin
Doctor of Science in Social Communications,
Senior Researcher, Director of Library
Maksymovych Scientific Library*
*Taras Shevchenko National University of Kyiv
Volodymyrska str., 60, Kyiv, Ukraine, 01033

Received date 07.07.2021
Accepted date 20.08.2021
Published date 25.08.2021

How to Cite: *Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S., Serbin, O. (2021). Devising a combined method for identifying near-duplicates in electronic scientific papers. Eastern-European Journal of Enterprise Technologies, 4 (4 (112)), 57–63. doi: https://doi.org/10.15587/1729-4061.2021.238318*

1. Introduction

The task of analyzing the content of electronic scientific papers for the identification of near-duplicates is relevant for professional scientific publications, specialized academic councils for the presentation of dissertations, and the scientific community in general. Improving the methods for identifying near-duplicates of scientific papers is an important tool for preventing abuse and plagiarism in the field of higher education and ensures academic integrity. However, the problem of identifying near-duplicates is not easy, since electronic scientific works can contain data of different types: texts, mathematical formulas, tables, schemes and diagrams, pictures, numerical data, etc. For the qualitative identification of near-duplicates, the data of all types must be analyzed for similarities using various methods that are best suited for analysis. That is why there is a problem of devising a combined method for identifying near-duplicates in scientific papers, taking into consideration data of various types.

An electronic scientific paper is a description of scientific research published electronically on the Internet, which meets

the key requirements for the design of scientific papers. An electronic scientific paper includes an analysis of a scientific problem or a task, research methods, results, and conclusions. An important prerequisite for a high-quality electronic scientific paper is to arrange its peer review before publication on the network. In this process, the key role is played by the identification of near-duplicates, the existence of which in an electronic scientific paper may indicate borrowing of third-party information without its citation, which is a copyright infringement. In general, the existence of near-duplicates in a scientific paper without appropriate citation means that a paper cannot be admitted to scientific review.

The task of identification of near-duplicates is of particular relevance for specialized academic councils and experts admitting a dissertation to the defense. A large number of scientific materials that must be presented by the author of the dissertation research may contain near-duplicates without citations to other studies in the area of a dissertation. Such abuse should be detected at the stage of consideration of a dissertation by experts and returned to the authors with a reasoned indication of the essence of violations. Thus, the task of devising a combined method for identifying near-duplicates in electronic scientific papers is relevant for education and science in general and ensures academic goodness and quality of scientific materials.

2. Literature review and problem statement

Analysis of sources should be divided into several components, taking into consideration the type of data to which the problem of identification of near-duplicates is applied. To solve the problem of identifying near-duplicates in images, paper [1] gives a description of a scheme for comparing an image with the images that are included in the corresponding databases to find a similarity between them. This scheme has acquired further development in paper [2]. The main drawback of the proposed schemes is great computational complexity. As a result, the search for near-duplicates may take an unacceptably long time. Another strategy for comparing images is to analyze each pixel separately. Paper [3] described this approach. However, its drawback is its dependence on the image size. In article [4], it is proposed to use comparisons of local image sections to identify near-duplicates. The effectiveness of this approach was shown in article [5]. The use of k-bit hash codes for this task was described in [6]. Paper [7] described a quick method for identifying near-duplicates in images that uses an intelligent method of analyzing similarities with the images published on the Internet. Paper [8] offers an improved algorithm for detecting near-duplicates of images, which uses the general feature of the color histogram, taking into consideration local complexity based on the calculation of entropy. This modification increases the accuracy of recognizing near-duplicates of images without a significant increase in calculations.

Analysis of text information similarity is often used to identify duplication in Web documents. In particular, paper [9] describes the problem of identifying near-duplicates for ranking web pages. Article [10] uses statistical analysis to avoid spam on the Internet. Article [11] describes the problem of making abstracts for digital libraries, which eliminates duplication of information. The described methods in the corresponding representation can be effectively used to identify near-duplicates of electronic scientific papers. Paper [12] describes a conceptual scheme for identifying near-duplicates in electronic documents. In particular, the method for identifying near-duplicates in tables based on locally sensitive hashing is described in research [13]. In this work, the concept of search is based only on the calculation of similarity. Paper [14] describes the method of n-gram analysis for identifying near-duplicates, which is a modern approach to establish similarities in text data. To identify the subjects of research of authors, in particular, of dissertation research, it is necessary to apply latent-semantic analysis. It was shown that the field of the study directly depends on the content of a scientific paper, which affects the quality and speed of identification of near-duplicates [15]. Paper [16] proposes the clustering method based on representative merge (Merge-Filter-RC) to detect near-duplicates in one or more data sources. Paper [17] describes the use of the TDW matrix. Each element of the matrix represents the frequency of the term in a document multiplied by weight. The importance of a term is assumed to vary depending on the location on the page. The authors make a predefined list of weights based on the HTML tags in which the term appears and assign weights from that list. The TDW matrix is used in paper [18] to identify near-duplicates, taking into consideration filtering of documents by the number of sentences. Article [19] describes the methods for measuring the similarity of a text to identify near-duplicates that can be used in borrowing detection systems in electronic scientific papers. However, scientific papers contain the content of different types: text, mathematical formulas, tables, images, etc. For qualitative identification of near-duplicates, the data of all types must be indexed and checked for borrowing.

3. The aim and objectives of the study

The purpose of this study is to develop a combined method for identifying near-duplicates in electronic scientific papers, taking into consideration the data of various types. This will make it possible to qualitatively identify near-duplicates and detect possible abuses and plagiarism in electronic scientific papers as widely as possible.

To achieve the aim, the following tasks were set:

– to analyze the methods for identifying near-duplicates in electronic scientific papers that contain the content of the same type, for example, text data, mathematical formulas, numerical data and consider using these methods to devise a combined method for identifying near-duplicates in electronic scientific papers;

– to develop an algorithm for the implementation of the combined method for identification of near-duplicates, which combines methods for identifying near-duplicates of data of various types;

– to verify the combined method for identification of near-duplicates in electronic scientific papers.

4. Materials and methods of research

The research used the methods of analysis, processing, and storing big data to identify similarities in databases using hash functions. The method of locally sensitive hashing and the method of comparison with the sample for identification of near-duplicates in electronic scientific papers were used. The method of comparative analysis was used to prove the effectiveness of the combined method for identifying near-duplicates.

An experiment on the identification of near-duplicates of electronic scientific works based on commercial services: Turnitin, Advego, and the system devised by the authors was carried out. To canonize the text of scientific papers in the developed system, a dictionary of stop words was created and a dictionary of the Ukrainian language was used. However, the functionality of the system makes it possible to use dictionaries of another language to canonize a text.

5. Results of devising the combined method for identification of near-duplicates in electronic scientific papers

5. 1. Analysis of the methods for identifying near-duplicates in electronic scientific papers containing the content of the same type

Let us assume that T is the input electronic scientific paper and $\{T_1,T_2,..., T_p\}$ are electronic scientific papers that were indexed and stored in a database, p is the number of scientific papers in a database. It is required to find such a set of scientific papers $\tilde{T}=\{T_j\}_{j\in\{1,2,...p\}}$, for which distance Φ between the papers from this set to the input scientific paper T does not exceed threshold value α, in other words,

$$\Phi\left(T,\overline{T}\right)\le\alpha,\ \forall\overline{T}\in\tilde{T}.$$

The method of identification of near-duplicates in the content of scientific papers of the text type. Let the text of the paper be assigned as a sequence of words:

$$W=\left\{w_1,w_2,\ldots,w_q\right\}, \tag{1}$$

where W is the specific text of an electronic scientific, w is the words of the electronic paper, q is the number of words.

Word w_i, $i=\overline{1,q}$ of arbitrary text W can be assigned as a sequence of letters:

$$w_i=\left\{l_1^i,l_2^i,\ldots,l_{u_i}^i\right\}, \tag{2}$$

where l_j^i, $i=\overline{1,q}$, $j=\overline{1,u_i}$ are the letters of the fixed alphabet, $l_j^i\in A$, u_i are the number of letters in word w_i.

We do not take all the characters that are not letters into consideration. Such text of paper W will be called a unigram. We will construct a new unigram, excluding from consideration the words that do not have a content load, primarily conjunctions. First, we will form a set of the following words:

$$Z=\left\{\text{«and»},\text{«but»},\text{«because»},\text{«in order to»},\text{«etc.»},\ldots\right\}.$$

It should be noted those compound conjunctions, such as: in order to; because; due to the fact that, etc. are written separately. Then a new unigram will include all the words of a scientific paper that do not belong to set Z, $\overline{W}=\{w_i\}_{w_i\notin Z}$. Assume that the capacity of such unigram $\overline{q}=card\left(\overline{W}\right)$, $\overline{q}\le q$.

Using the method of sliding window based on $\overline{W}$, construct a unigram of length r:

$$\overline{W_j}=\left\{w_i\right\}_{\substack{i=j,\\ w_i\notin Z}}^{r+j-1},\ j=\overline{1,\overline{q}-r+1}, \tag{3}$$

where r is the size of a sliding window.

Based on the method of locally sensitive hashing, represent a set of unigrams $\{\overline{W_j}\}_{j=1}^{\overline{q}-r+1}$ in the form of a set of bit series $\{I(\overline{W_j})\}_{j=1}^{\overline{q}-r+1}$, where $I(\overline{W_j})$ is the element of index that at the same time determines the sequence $\overline{W_k}$, $j=\overline{1,\overline{q}-r+1}$,

$$I\left(\overline{W_j}\right)=\left\{\delta_{jk}\right\}_{k=1}^{t},\ \delta_{jk}\in\{0,1\}, \tag{4}$$

t is the number of bits in a series.

Assume that $I(\overline{W_j^A})$, $A=\overline{1,p}$ are the elements of the index for electronic scientific papers that are stored in the database, p is the number of scientific papers in the database, in this case:

$$I\left(\overline{W_j^c}\right)=\left\{\delta_{jk}^c\right\}_{k=1}^{t},\ \delta_{jk}^c\in\{0,1\},\ A=\overline{1,p},\ j=\overline{1,\overline{q}-r+1}. \tag{5}$$

The Hamming distance between the elements of the index of an incoming electronic scientific paper stored in the database is:

$$\Phi_H^c=\frac{1}{t}\sum_{k=1}^{t}\left|\delta_{jk}^c-\delta_{jk}\right|, \tag{6}$$

where Φ_H^c is the Hamming distance between the elements of the index of an incoming scientific paper and the paper from a database with number c, $c=\overline{1,p}$.

If $\Phi_H^c\le\alpha_H$, $\alpha_H\in[0,1]$, we can consider that the element of the index of an incoming scientific paper is close to the corresponding element of the index of the paper with the number c, that is, the incoming electronic scientific paper contains an near-duplicate.

The method for identification of near-duplicates in the content of scientific papers of numerical type. Let us assume that the text of the paper contains the numeric values that are assigned in the form of:

$$N=\left\{n_1,n_2,\ldots,n_v\right\}, \tag{7}$$

where N is the set of numerical values of an electronic scientific paper, v is the quantity of numbers.

Using the method of sliding window based on N, the sub-sequences of length r were constructed:

$$N_j=\left\{n_i\right\}_{i=j}^{r+j-1},\ j=\overline{1,v-r+1}, \tag{8}$$

where r is the size of the sliding window or the length of a sub-sequence.

Let us assume that electronic scientific papers stored in the database contain numerical values that are represented as sub-sequences:

$$N_j^c=\left\{n_i^c\right\}_{i=j}^{r+j-1},\ j=\overline{1,v-r+1},\ c=\overline{1,p}, \tag{9}$$

where n_i^c is the i-th numerical value of the electronic scientific paper with number c.

Then to calculate the distance between the numerical components of an input scientific paper and the numerical components of the papers stored in the database, it is possible to use one of the metric distances, for example, the Euclidean distance:

$$\Phi_E^c=\sqrt{\sum_{j=1}^{r+j-1}\left(n_j-n_j^c\right)}, \tag{10}$$

where Φ_E^c is the Euclidean distance between numerical components of electronic scientific papers.

The method for identification of near-duplicates in the content of electronic scientific papers, consisting of mathematical formulas.

The peculiarity of scientific papers is that most of them include mathematical formulas that have a complex structure of symbolic designations. The difficulty of analyzing mathematical formulas is that it is necessary to analyze not only the graphical image of the formula but also the text interpretation of a formula, taking into consideration possible changes in designations, numeric values, the shape of brackets, etc.

Mathematical formulas are mostly created using special formula editors. Formula editors can be created from:

– using a special markup language, for example, TeX, MathML;

– using graphic interface: MathType, KFormula, MathCastmula, MathCast, etc.;

– using embedded components: Math Expression Editor Light;

– using symbolic calculations: Mathematica, MatLab, etc.

However, markup languages and formula editors such as MathType are used to type the text of scientific papers, in particular, of dissertations. In particular, according to TEX rules, formulas are separated by @@ characters, which allows recognizing these characters. In electronic scientific papers, separate mathematical symbols or small formulas that are placed in the text are separated by $...$ characters, formulas that are types in separate lines – by $$...$$ characters. The letters of the Latin, Greek alphabet («\alpha», «\beta»), etc. are used to designate variables. Special marks (e.g. «+», «–», «=», «\neq' (≠)», «\sum» (∑), etc.) are used to designate mathematical operations, marks (e.g. «\sin», «\min», «\ln», etc.) are used for simple mathematical functions. Understanding the marks of special markup language makes it possible to compare the formulas of electronic scientific papers. In this case, possible changes in designations are taken into account. That is, the comparison should be carried out according to the structure of the formula, and not according to designations.

If mathematical formulas are displayed as a MathType object, it is convenient to use the sampling method to compare a formula. First, samples are compared, and then the names of variables. It should be borne in mind that certain mathematical, physical, and chemical formulas can have stable generally accepted designations and they will not be near-duplicates. It will be appropriate to create a designation dictionary for documents, taking into consideration the description of designation in formulas, which are often specified immediately after the formulas.

The variants of representation of text objects with formulas were considered:

1. Mathematical formulas are saved as MathType objects.

2. Mathematical formulas are saved as graphic objects (.jpg).

3. Mathematical formulas are saved in the format (.pdf).

In the first case, the comparison algorithm by the sample is used:

1. Formation of samples of formulas of an incoming electronic scientific paper.

2. Checking the data of a sample with the samples of formulas of scientific papers included in the database.

3. For samples with the same structure, the name of the variables is checked.

4. In the case of a full or partial coincidence, we can talk about an near-duplicate detected.

The MathType Editor contains a translator of mathematical expressions into MathML format, which simplifies the problem of comparison with the sample.

In the second case, it is necessary to use the OCR technology of optical character recognition with structural analysis. An important task, in this case, is to convert graphic images of formulas into MathML and to apply the already described scheme.

In the third case, there are difficulties with converting formula images into the .pdf format and bringing them to MathML. However, in general, the concept of transformations coincides with the second case.

The described scheme makes it possible to analyze similarities and identify duplication of formulas both using samples of analyzed formula objects and using the conversion of mathematical formulas in different formats into the MathML mathematical marking language.

The method for identifying near-duplicates in schemes and diagrams.

Let us assume that diagram or scheme D were detected in an incoming electronic scientific paper. Compare it with the diagrams of electronic scientific papers stored in the database $\{D_1, D_2,..., D_p\}$. The comparison algorithm takes the form:

1. Form code representations of diagrams and schemes D, $D_1, D_2, ..., D_p$.

2. Comparison of code representations with each other.

3. If a partial or complete match of code representations is detected, a text description of schemes and diagrams is compared with each other.

4. Making a decision on the existence of near-duplicates in schemes and diagrams.

In the case of comparing images in electronic scientific papers, which are considered as graphic objects, it is necessary to take into consideration the types and formats of these images. Among the methods for processing graphic information, it is possible to separate two main directions: finding key points of an image and using locally sensitive hashing for individual pixels of an image.

A separate task of identifying near-duplicates in electronic scientific papers is to compare tables. In case of hiding borrowings, the original table can be easily modified: changing the places of rows and columns, changing headings and field names. In addition, when it comes to unfair borrowings, these tables can be significantly changed. For example, if the original table contains the results of a numerical experiment, these numerical data can be deliberately changed in the borrowing. The task of finding near-duplicates in tables is the process of identification of such tables that are most similar to each other. The similarity, in this case, is expressed by certain functional Φ, which assigns the distance between the tables. If this distance does not exceed the threshold value α, the tables are treated as similar, and therefore, there are near-duplicates in the data of these tables. Tables of electronic scientific papers can contain text, numerical data, formulas, data such as a date, etc. That is why the concept of finding near-duplicates in tables is generally similar to the concept of searching for near-duplicates in electronic documents in general [13].

5. 2. Algorithm for implementing the construction of a combined method for identifying near-duplicates of various types of data

The algorithm for the implementation of the combined method for identification of near-duplicates in electronic scientific papers consists of the following stages:

1. To separate images, including schemes and diagrams, numerical data, tables, formulas, and text from an input electronic scientific paper.

2. According to formulas (1) to (6) analyze a text for near-duplicates.

3. According to formulas (7) to (10), to analyze numerical data for the existence of near-duplicates with electronic scientific papers stored in the database.

4. To analyze mathematical formulas: to form samples and compare them, compare designations in case of similarity of samples, taking into consideration commonly used designations. This is done using the method for identifying near-duplicates in the content of electronic scientific papers, consisting of mathematical formulas.

5. To index tables, to select from them separately numeric and text data and separately analyze for near-duplicates. The method for identification of near-duplicates in the tables is described in detail in paper [13].

6. Compare schemes, diagrams, and other images existing in an electronic scientific paper by identifying near-duplicates in schemes and diagrams.

7. If near-duplicates without reference to electronic scientific works were found in the content, the corresponding incoming electronic scientific paper is sent for examination. The examination establishes whether an electronic scientific paper contains borrowings without a reference and can be qualified as plagiarism. The existence of near-duplicates for each of the specified types is determined by the appropriate method.

The algorithm made it possible to devise a test system that implements the method for the identification of near-duplicates in electronic scientific papers presented in the HTML format. In this case, the image was stored directly in the text using the BASE64 encoding. The system contains the functions of import of electronic scientific papers, which consists of the following stages:

1. Upload a file to the server.

2. Specify the format of the uploaded file. The system supports formats.

3. If the format of the uploaded document is different from HTML, the system determines the converter for this format. The system supports the conversion of PDF, DOCX, ODT, RTF, and TXT formats.

4. Convert the uploaded file to HTML and clean it from styles and scripts.

5. Save the received document in the database.

If the language of an electronic scientific paper differs from the Ukrainian language, the system performs automatic translation using the Google Cloud Translation API. In this case, the text is fragmented into short passages of one or more sentences up to 500 characters in one fragment.

For each document, the text is canonized and indexed using locally sensitive hashing. After preliminary processing and canonization of the text data of tables in an electronic scientific paper, fragmentation, and creation of the table index take place using locally sensitive hashing.

Image processing includes fragmentation at specific key points and determining the own rotation angle of each fragment, based on which the perceptual hash is constructed. To determine the own rotation angle, not the entire image rotates, but rather each sector separately, in which the average position color is calculated. This limitation is caused by the high computational complexity of the rotation operation and makes it possible to index images much faster, but in this case, some information is lost.

5. 3. Verification of the combined method for identification of near-duplicates in electronic scientific papers

35 electronic scientific papers of the authors were selected to verify the method. Electronic scientific papers were divided into three groups in the areas of scientific research: scientometrics, antiplagiarism, monitoring of environmental pollution. Publications were checked using the Turnitin, Advego system and using the system that implements the method described in the article. All authors' publications were indexed into the database of the system, but during verification, the publication that was being checked was excluded from the database. This is necessary for the correctness of the check. The database of the system included publications, full texts of which were obtained from the Vernadsky National Library of Ukraine [20].

The comparison of the uniqueness of scientific papers was made by the Turnitin, Advego services, and using a system that implements the combined method for identifying near-duplicates for each group of publications corresponding to the relevant direction of research. Since the Advego service has a free verification limit of 3,000 characters, the document was divided into parts. The general uniqueness of an electronic scientific paper was calculated taking into consideration the degree of uniqueness of each of the parts. The test system was checked in two versions: checking a text only (option 1), checking a text and other objects (images, formulas, tables, etc.) (option 2).

Table 1 shows average degrees of uniqueness obtained as a result of checking the test set of electronic scientific papers in three groups by the direction of scientific research (group 1 – scientometrics, group 2 – antiplagiarism, group 3 – environmental monitoring).

According to the results of identification of near-duplicates in 35 electronic scientific papers, according to the devised system of using the combined method according to option 1, it was possible to detect by 3.4 % more borrowings than by the method of identification of text borrowings (option 2).

Table 1

Comparison of the combined method for identifying near-duplicates with the methods used by Turnitin and Advego services

Groups of publications	Number of publications	Average volume (number of words)	Average degree of uniqueness (Turnitin)	Average degree of uniqueness (Advego)	Average degree of uniqueness (Test system, option 1)	Average degree of uniqueness (Test system, option 2)
Group 1	19	2757	78	94	81.8	77.8
Group 2	11	2598	73	92	78.1	75.7
Group 3	5	2390	77	100	79.6	76.2

6. Discussion of results of studying the methods for identifying near-duplicates in electronic scientific papers

Since electronic scientific papers contain the content of different types: text numerical data, formulas, images, in particular, schemes and diagrams, the methods for the identification of near-duplicates in the content of each of the data types were used separately to implement the combined method. For text data, the method of locally sensitive hashing with finding the Hamming distance between the elements of electronic scientific papers indexes was used. For numerical data, the method for constructing a sub-sequence for each scientific paper with determining the proximity between the vectors consisting of the numbers of these sub-sequences was used. To compare mathematical formulas, the method for comparing samples of formulas was used. To identify near-duplicates in graphic information, two directions were used: finding key points in the image and applying locally sensitive hashing for separate pixels in the image. Each of these methods is effective enough to identify near-duplicates in the content of the same type, respectively, these methods can be used to implement the combined method.

The devised algorithm of the combined method for the identification of near-duplicates made it possible to implement the test system. Using this system, this method was verified on a set of 35 electronic scientific papers of authors containing the content of various types. The results of the verification are included in Table 1. The degree of borrowing for Table 1 is explained by the fact that the authors are working on scientific projects, in which the articles that were checked, are in the cycle of scientific publications. Accordingly, each subsequent publication relies on the results obtained in the previous one and contains an abridged description of the results obtained in the past.

The Advego service searches the Internet and does not find some sources that were not indexed by search engines. The system in option 2 finds more borrowings mostly through identified near-duplicates in mathematical formulas. Accordingly, the percentage of uniqueness decreases.

Unlike the methods for identifying near-duplicates in the content of the same type (for example, text data), the combined method for identification determines borrowings in the content of different types. This is important because electronic scientific papers usually contain data of different types: text, formulas, tables, images, etc. The construction of the combined method becomes possible due to the use of a unified approach to the indexing of various types of data, representing the content of electronic scientific papers.

To apply the developed method in practice, it is necessary to index a large volume of electronic scientific papers from various fields. There is a dependence of the results of the combined method for the identification of near-duplicates on the database of indexed scientific papers. The indexing process has computational complexity, which is a certain limitation on the application of this method in practice.

The components of the combined method can include other methods that are not considered in this article. This requires separate research.

7. Conclusions

1. The methods for identification of near-duplicates in electronic scientific papers containing content of the same type, for example, text data, mathematical formulas, numerical data, tables, schemes, diagrams, and other images were analyzed. The degree of proximity of fragments of an electronic scientific paper to the papers included in the scientific database is calculated for the text data by calculating the Hamming distance between the elements of indices, which are formed by the method of locally sensitive hashing. To compare numerical data, numerical sub-sequences are formed and Euclidean distance between the vectors that consist of these sub-sequences is calculated. To compare mathematical formulas, the method of samples comparing is used. For graphic information, the method of finding key image points and the method of locally sensitive hashing for image pixels are used. If an image is a diagram or a chart, the text of an image is separately selected and analyzed. The implementation of these methods of identification of near-duplicates made it possible to rationally organize the construction of the combined method.

2. The algorithm of implementation of the combined method for identification of near-duplicates, which combines the methods for identification of near-duplicates of various types of data, was devised. The algorithm made it possible to develop the test system that implements the method for the identification of near-duplicates in electronic scientific papers presented in HTML format.

3. To verify this method, we selected electronic scientific works of authors, which were divided into three groups according to the direction of research. The papers were analyzed for borrowings in the Turnitin system, using the Advego service and using the system that implements the combined method. It was found that the combined method allows identifying near-duplicates both in text information and comparing other objects of scientific papers with the database of scientific papers. In particular, according to the results of identification of near-duplicates in 35 electronic scientific papers, using the developed system, the use of the combined method made it possible to detect 3.4 % more borrowings than the method of identification of text borrowings.

References

1. Wu, X., Ngo, C.-W., Hauptmann, A. G. (2008). Multimodal News Story Clustering With Pairwise Visual Near-Duplicate Constraint. IEEE Transactions on Multimedia, 10 (2), 188–199. doi: https://doi.org/10.1109/tmm.2007.911778
2. Chang, E. Y., Wang, J. Z., Li, C., Wiederhold, G. (1998). RIME: A replicated image detector for the World Wide Web. Proceedings of SPIE - The International Society for Optical Engineering, 3527, 58–67. doi: https://doi.org/10.1117/12.325852
3. Liu, G.-H., Yang, J.-Y. (2013). Content-based image retrieval using color difference histogram. Pattern Recognition, 46 (1), 188–198. doi: https://doi.org/10.1016/j.patcog.2012.06.001
4. Mikolajczyk, K., Schmid, C. (2005). A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (10), 1615–1630. doi: https://doi.org/10.1109/tpami.2005.188

5. Ke, Y., Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. doi: https://doi.org/10.1109/cvpr.2004.1315206
6. Zou, F., Feng, H., Ling, H., Liu, C., Yan, L., Li, P., Li, D. (2013). Nonnegative sparse coding induced hashing for image copy detection. Neurocomputing, 105, 81–89. doi: https://doi.org/10.1016/j.neucom.2012.06.042
7. Gadeski, E., Le Borgne, H., Popescu, A. (2016). Fast and robust duplicate image detection on the web. Multimedia Tools and Applications, 76 (9), 11839–11858. doi: https://doi.org/10.1007/s11042-016-3619-4
8. Li, Y. (2021). A Fast Algorithm for Near-Duplicate Image Detection. 2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID). doi: https://doi.org/10.1109/aiid51893.2021.9456496
9. Yi, L., Liu, B., Li, X. (2003). Eliminating noisy information in Web pages for data mining. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '03. doi: https://doi.org/10.1145/956750.956785
10. Fetterly, D., Manasse, M., Najork, M. (2004). Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. Proceedings of the 7th International Workshop on the Web and Databases Colocated with ACM SIGMOD/PODS 2004 – WebDB '04. doi: https://doi.org/10.1145/1017074.1017077
11. Chang, H.-C., Wang, J.-H. (2007). Organizing News Archives by Near-Duplicate Copy Detection in Digital Libraries. Lecture Notes in Computer Science, 410–419. doi: https://doi.org/10.1007/978-3-540-77094-7_52
12. 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
13. 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
14. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., 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. doi: https://doi.org/10.15587/1729-4061.2019.175139
15. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S. (2020). 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, 4 (4 (106)), 21–28. doi: https://doi.org/10.15587/1729-4061.2020.209886
16. Fellah, A. (2021). All-Three: Near-optimal and domain-independent algorithms for near-duplicate detection. Array, 11, 100070. doi: https://doi.org/10.1016/j.array.2021.100070
17. Mathew, M., Das, S. N., Lakshmi Narayanan, T. R., Vijayaraghavan, P. K. (2011). A novel approach for near-duplicate detection of web pages using TDW matrix. International Journal of Computer Applications, 19 (7), 16–21. doi: https://doi.org/10.5120/2374-3128
18. Arun, P., Sumesh, M. (2015). Near-duplicate web page detection by enhanced TDW and simHash technique. 2015 International Conference on Computing and Network Communications (CoCoNet), 765–770. doi: https://doi.org/10.1109/coconet.2015.7411276
19. Mishra, A. R., Panchal, V. K., Kumar, P. (2020). Similarity Search based on Text Embedding Model for detection of Near Duplicates. International Journal of Grid and Distributed Computing, 13 (2), 1871–1881. Available at: http://sersc.org/journals/index.php/IJGDC/article/view/35004/19401
20. National Library of Ukraine named after VI Vernadsky. Available at: http://nbuv.gov.ua/