ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)
Наука та прогрес транспорту, 2023, № 2 (102)

UDC 347.78:82

V. I. SHYNKARENKO[1*], I. M. DEMIDOVICH[2*], O. S. KUROPIATNYK [3*]

[1*]Dep. «Computer and Information Technologies», Ukrainian State University of Science and Technologies, Lazariana str. 2, Dnipro, Ukraine, 49010, tel. +38 (056) 373 15 52, e-mail shinkarenko_vi@ua.fm, ORCID 0000-0001-8738-7225
[2*]Dep. «Computer and Information Technologies», Ukrainian State University of Science and Technologies, Lazariana str. 2, Dnipro, Ukraine, 49010, tel. +38 (056) 373 15 52, e-mail 2019demidovichinn@gmail.com, ORCID 0000-0002-3644-184X
[3*]Dep. «Computer and Information Technologies», Ukrainian State University of Science and Technologies, Lazariana str. 2, Dnipro, Ukraine, 49010, tel. +38 (056) 373 15 52, e-mail olena.kuropiatnyk@gmail.com, ORCID 0000-0003-2286-884X

A Dual Approach to Establishing the Authority of Technical Natural Language Texts and Their Components

Purpose. The study is aimed at testing the hypothesis that it is possible to determine plagiarism by methods of establishing the authorship of a text without using a text bank and their direct comparison. **Methodology.** Constructive and productive models of the processes of establishing the authorship of technical texts for two methods have been developed. The first method is based on the formation of a text model in the form of a set of formal substitution rules with probabilistic weights (as in stochastic formal grammars), which reflects the syntactic features and patterns of text formation by the author. The degree of similarity between the text under study and another text is determined by comparing their models. The second method is a classical approach to detecting borrowings (plagiarism) by directly comparing the text under study with an existing text bank, highlighting repeated text fragments, and determining the degree of originality. Experiments were conducted to establish the correlation between the results of these two methods. The experimental base consisted of 509 text sections of theses of students majoring in «Software Engineering». **Findings.** Experimental studies have made it possible to establish a high correlation between the results of the two methods. Correlation coefficients in the range of 0.75...1.0 and with an average value of 0.88 were obtained provided that borrowings are taken into account for text fragments of at least five words in length. **Originality.** For the first time, the authors have identified the possibilities and proposed methods for indirect plagiarism detection without using a large text bank. The essence of the model is to formalize the representation of the author's sentence syntax by a set of substitution rules with probabilistic weights. **Practical value.** Based on the results obtained, the possibilities for detecting borrowings have been expanded and the effectiveness of the corresponding methods has been increased. Recommendations on the parameters of classical methods for detecting borrowings have been obtained, in particular, it is recommended to take into account text fragments of at least five words in length as a rational parameter when using borrowing detection systems. The possibilities of text authorship detection methods tested on fiction texts are extended to technical texts.

Keywords: natural language texts; authorship determination; statistical analysis; classification; correlation coefficient; constructive-productive modeling; constructivism; formal grammars; graphs

Introduction

The problem of identifying similarities and differences in the various authors' texts is still relevant due to the difficulties of identifying commonalities that are not a direct coincidence of the text. A special difficulty is working with specific characteristics of a certain language, which significantly complicates the task and makes it impossible to create a unified toolkit.

Currently, approaches from the theory of pattern recognition, mathematical statistics and probability theory, algorithms of neural networks and cluster analysis, and many others are used for text attribution. However, all such methods do not have sufficient efficiency and cannot work with texts of different languages and topics, and also do not work with the stylistic features of the author to a sufficient extent.

Consider two tasks for processing natural language texts:

– the task of borrowings identifying (establishing the authorship of individual parts of texts – phrases, sentences, paragraphs, the text as a whole). There is a text, and it is necessary to highlight those parts of it that are already found in earlier texts by other authors and establish the degree of the text originality;

doi: https://doi.org/10.15802/stp2023/288958

– the task of establishing the text authorship according to the style and other features of the author's text.

To solve the first problem (the problem of borrowings identifying), it is necessary to have a complete bank of texts to compare with. Taking into account their huge number, and the presence of various storages and storage methods (in particular, file formats), there is a certain probability of positively erroneous decisions, i.e. part of the borrowings will not be detected.

For the second task (the task of establishing syntactic similarity), it is necessary to have a certain number of the author's texts, at least one text of sufficient length. At the same time, no specific borrowings are identified, but conclusions are drawn about the text as a whole. Previously, the authors of the work should pay attention to the specificity of the speech by a specific author [15, 26].

In this paper, we study the correlation dependence between the results of solving these two problems. For this, the approach of constructive-synthesizing modeling proposed by the authors for solving both problems is applied and the corresponding methods are developed.

Both tasks are solved using software tools that implement the constructors presented in this article. They are multi-parametric and solve the problem with varying degrees of accuracy. The expert must make the final decision. For an expert, the results of checking by one of the methods (solution of one problem) may be sufficient. For a more objective expert decision, both of these methods can be used.

Related works

One of the problems considered in this paper is the identification of borrowings. Borrowing is a fairly common problem in academic fields, including scientific articles, publications, inventions, etc. [1]. Plagiarism comes in many varieties; for example, self-plagiarism (publishing the same or very similar articles in several journals) or using the texts of other authors.

This phenomenon can be observed in both academic and non-academic environments. Academic plagiarism is one of the most serious forms of academic misconduct and negatively affects the educational institution and its employees. Research articles containing, among other things, plagiarism, interfere with the scientific process [25]. The existence of plagiarism can have serious consequences. Plagiarism of research articles can significantly affect the work of specialists in various fields, for example, plagiarism in the medical field can threaten the safety of patients [25]. In addition, plagiarism wastes scientific resources. Even detecting, investigating and punishing plagiarized research articles requires a lot of effort by academics, institutions and funders [25].

There are many methods of working with borrowings and their detection. All of them can be grouped: lexical detection methods [4] (working only with symbols or their sequence of a certain length [9] in a document or even words [5]); detection methods based on lemmas [8] and syntax (working with the syntactic structure of a sentence, i.e. parts of speech) [10], grammars [15, 27], detection methods based on semantics [13, 20] and comparing a certain sequence of words [21] or sentences [6]; detection methods based on ideas and contents go beyond the analysis of the text in the document, for example: the mathematical component [14], citations [22] and images in it [18].

Checking a suspected document for plagiarism manually is an extremely difficult and time-consuming process for different source documents [1]. Therefore, the use of computer systems is appropriate. The plagiarism detection tools that have been proposed so far are capable of detecting different types of plagiarism; however, the detection of plagiarism in the text depends on experts [19].

In Ukraine and other countries, means of detecting plagiarism and borrowing have been introduced in the academic environment and universities. However, even with sufficient efficiency and credibility of the work, it is not possible to ensure coverage of all sources of plagiarism due to the constant increase in their number and free access to them on the Internet.

A hypothesis is put forward about the possibility of identifying borrowings by methods related to establishing the authorship of texts based on the analysis of the author's existing text.

The second problem, which is the subject of this research, is establishing the authorship of the text. Accurate and reliable authorship establishment requires the use of a certain texts' corpus by different authors, which will allow establishing a style characteristic of them and subsequently us-

doi: https://doi.org/10.15802/stp2023/288958

ing this to establish the authorship of other texts.

Methods for solving this problem belong to the same groups as methods for identifying borrowings. Due to the formalization of the text, they have a wide range of applications for different languages in the world. An example of the methods and approaches range to solve the problem is the use of neural networks for Ukrainian texts [17], a genetic algorithm for working with Turkish texts [10], establishing the authorship of ancient texts in Hindi [22], using the peculiarities of the parts of the language and different stylistics usage [7], as well as features of working with small texts [2] and even text messages [12].

However, none of the methods or their combination still gives 100% accuracy in determining the authorship.

Purpose

The research aiming to test the hypothesis of the determining plagiarism possibilities by establishing the text authorship methods without using a texts bank and their direct comparison.

Methodology

The processes of texts authorship identifying using the constructive-production modeling. To determine the authorship of texts, a constructive-production approach to modeling the sentence structure is used. This process consists of the sequential work of the following constructors: the constructor-converter of natural language text into tagged text, the constructor-converter of tagged text into a set of formal substitution rules with a probability measure, and the constructor-measurer of the two texts similarity degree. They are described in more detail in the authors' previous article [23].

Constructive and synthesizing modeling of borrowing detection processes. The graphic representation of the texts constructor. To speed up the comparison of the text, it is suggested to use the designer-converter of the natural language text into a graph. The idea of the constructor's work is to create a graph structure based on the text, which contains all the chains that have the input text and does not contain extraneous ones.

The constructor has the form:

$$C =< M, \Sigma, \Lambda >_S \mapsto C_g =< M_g, \Sigma_g, \Lambda_g >,$$

where M_g – is an extensible medium that includes sets of graph constructions, language constructions (words, sentences, etc.) and their elements, Σ_g – is a set of operations and relations on the elements M_g, Λ_g – is a set of CIS statements.

Claims about the carrier. The carrier includes multiple terminal and non-terminal elements $M_g \supset T_g \bigcup N_g$. Terminals are language constructions constructed by ${}_A C_T$ the constructor and their components (T_T), as well as graph constructions and their components $T_g = \bar{\Omega} \bigcup \Omega_g \bigcup T_T \bigcup V \bigcup E$, Ω_g – is a set of graph constructions, V, E – sets of vertices and arcs with their attributes.

The vertex has the attributes $\bar{w}_v = \langle id, content \rangle$, id – identifier, accepts integer values, *content* – part of the text structure. Attributes of the arc $\bar{w}_e = \langle id, routes, start, end \rangle$, id – identifier, takes integer values, *routes* – set of numbers of the paths in which the arc is included (indicates the order of traversal of the graph), *start*, *end* – vertices that are incident to the arc *e*.

Denote a loaded graph as ${}_{\bar{w}_g} G = \langle V, E \rangle$, $V = \{ {}_{\bar{w}_{vi}} v_i \}, E = \{ {}_{\bar{w}_{ej}} e_j \}$ – is the set of vertices and arcs loaded with attributes. Each set contains an empty element.

The graph has the following attributes $\bar{w}_g = \langle start\_v, last\_v, current\_v, amount\_l \rangle$, where $start\_v$ – is the starting vertex of the graph, $last\_v$ – is the last added vertex, $current\_v$ – is the current vertex when forming the graph, and $amount\_l$ – s the number of cycles that the starting vertex includes.

Operations statements. A substitution relation and several specified operations are used to construct the graph [25]:

- definition of an arc by incident vertices;
- finding a vertex with a content weight attribute equal to the given value;
- execution of *n* operations from the given list;
- calculation of a set power;
- addition of two numbers;
- union of graphs;
- partial and complete removal.

doi: https://doi.org/10.15802/stp2023/288958

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)
Наука та прогрес транспорту, 2023, № 2 (102)

The complete ontology of the graph constructor, as well as the specification of the graph construction rules, are presented in [26].

The goal of construction is to construct a graph structure that corresponds to a given text structure.

The graph constructor is limited by the text construction. The number of graphs depends on the number of different characters in the text.

Initial conditions for the construction of graphs: σ – a non-terminal from which the derivation begins.

Construction completion condition: the form does not contain non-terminals and each text construction element corresponds to a graph construction element.

Text comparison is performed as follows: two texts are input: in the form of a line and an ordered graphs set; a character-by-symbol comparison of the string text with the text in the graphical representation is performed. The graph is selected so that its starting vertex has a content attribute *content* equal to the current character in the text. After establishing this equality, traversal is performed in the graph in the order that corresponds to the order in the text-line. If the specified order cannot be found in the graph, then the transition to the next word in the line and another graph in the set is performed. The result of the comparison is the lines set – fragments that occur in both input texts. After that, the percentage of borrowings can be calculated as the ratio of the total found fragments length to the length of the input text string.

Let's formalize the specified processes using the constructor for comparison:

$$C =< M, \Sigma, \Lambda >_S \mapsto C_C(t, \{G_i\},$$

$$param) =< M_C, \Sigma_C, \Lambda_C >,$$

where M_C – is a carrier including sets of terminal (T_C) and non-terminal (N_C) symbols, Σ_C – is a set of operations and relations on the elements of M_C, Λ_C – is a set of CIS statements; t – text in the form of a string – a sequence of characters t_i (reformatted text in which the paragraph and line break symbols are replaced by spaces); $\{G_i\}$ – the text presented in the form of an ordered set of graphs, which is the result of the work of the C_g, *param* – the minimum number of words of the fragment, which is considered a borrowing.

The carrier claims. $M_C \supseteq T_C \bigcup N_C$, $T_g = \bar{\Omega} \bigcup \Omega_g \bigcup T_T \bigcup V \bigcup E \bigcup X_{\text{int}} \bigcup R$, where X_{int} is the integers' vectors set of indicating the beginning and end positions in the text fragments that are the same in the two texts, R – is a set of real numbers, which includes the percentage of borrowings in texts that are checked for originality.

Statement of operations. A substitution relation and several specified operations are used to compare text presented as a string and text in a graphical representation**:**

– $\bullet(G_f, t_i, \{G_i\})$ – is linking the symbol t_i of the string text with the corresponding graph $G_f \in \{G_i\} : content \leftarrow\!\lrcorner start\_vertex \leftarrow\!\lrcorner G_f = t_i$, where $a \leftarrow\!\lrcorner b$ – s an access operation to attribute a of entity b, $\{G_i\}$ – is text in the form of a set ordered by start_vertex graph sets;

– $\Delta(r, t_i, G_f, v_s, cr)$ – is a connecting the next character of the text t_i with the vertex of the graph G_f, while there is a transition from the v_s to the vertex v along the arc $_{wi}e_i \in E_f$, cr – is an integer, the route number in the graph G_f, r – is the operation result, $r = true$, if $\exists w_i \leftarrow\!\lrcorner e_i \in E_f : w_i = \langle id, routes, v_s, v \rangle$, while $v = start\_vertex \leftarrow\!\lrcorner G_f(cr+1) \in routes_i$, otherwise – $cr \in routes_i$, $routes_{i-1} \leftarrow\!\lrcorner e_{i-1} : end \leftarrow\!\lrcorner e_{i-1} = v$, $v_s, v \in V_f$, $G_f = V_f \bigcup E_f$, $weight \leftarrow\!\lrcorner v = t_i$ else – false;

– $+(c, a, b)$ – is an addition of two numbers a and b, $c = a + b$;

– $\times(i, f, s)$ – is an entry at the end of the list of boundaries of common fragments f boundaries s, i of the borrowed fragment, $s < i, s \geq 0, 0 < i < |t|$, $|t|$ – the length of the text submitted for comparison, in characters;

– $|t|$ – is a determination of the text length in characters *t*, presented as a string, $|t| \geq 0$;

– $=(a, b)$ assignment of entity a to value b;

– $\downarrow(q, t, f, param)$ – search q – the number of words in parts of the text t, the limits of which are

doi: https://doi.org/10.15802/stp2023/288958

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)
Наука та прогрес транспорту, 2023, № 2 (102)

determined f – a list of numbers, where each odd element is the position in the beginning borrowed fragment of the text, even – the end, *param* – the minimum number of words fragment, which is considered as a plagiarism;

– $/(c,a,b)$ – is a calculation of $c \in R, c \in [0;1]$ – is the fraction that makes up a from b;

– $<(c,a,b)$ – is a comparison of a,b, $c = true$, if *a* is less than *b*, otherwise $c = false$;

– $>(c,a,b)$ – is a comparison of a,b, $c = true$, if *a* is bigger *b*, otherwise $c = false$;

– $==(c,\ a,\ b)$ – is a comparison of a,b, $c = true$, if *a* is equal to *b*, else $c = false$;

– $\&(c,a,b)$ – is a logical «and» operation on operands *a* and *b*, *c* is the operation result.

The purpose of the construction is to establish the degree of the texts similarity by comparing the text in the form of a line with the text in the graphical representation built by the constructor C_g.

The initial construction conditions t – is the text in the form of a string in which borrowings are sought, $i = 0$ – is the number of the symbol in the text t, from which the comparison begins, $\{G_i\}$ – is the text in the form of an ordered set of graphs, σ, α – is an initial non-terminal (axioms).

Construction completion condition: getting a number from 0 to 1 that reflects the similarity of two texts represented as a string and a set of graphs.

The specification of the constructor for comparison:

$$C_C =< M_C, \Sigma_C, \Lambda_C >_K \mapsto C_C =< M_K, \Sigma_K, \Lambda_K >$$

where $M_K \supseteq M_C$, $\Sigma_K \supseteq \Sigma_C$, $\Lambda_K \supseteq \Lambda_C \bigcup \Psi_K$, $\Psi_K = \{\langle\{s_i\},\{g_i\}\rangle\}$ – is a set of production rules, s_i – is a substitution rule, g_i – are operations on attributes, $s_i = \langle s_{i,1}, s_{i,2}\rangle$, $g_i = \langle g_{i,1}, g_{i,2}\rangle$, $s_{i,1}, s_{i,2}$ – operations on text in the form of a line and a graph, respectively, operations $g_{i,1}, g_{i,2}$ are performed before and after the substitution rule s_i respectively.

Consider the rules $s_1 - s_5$, which allow forming a vector of positions (integers) f of borrowed fragments in the text t.

If the text string is not processed completely, that is, the current position is less than the length of the text $|t|$ (checked using $g_{1,1}, g_{1,2}$), the connection of the current character of the line with the graph G_f in the set $\{G_i\}$ is performed

$$s_{1,1} = \langle \alpha\ _{\tau 1} \rightarrow t_i \alpha \rangle, \hookleftarrow$$

$$s_{1,2} = \langle \sigma\ _{\tau 1} \rightarrow \bullet(G_f, t_i, \{G_i\})\beta \rangle,$$

$$g_{1,1} = \langle <(\tau 1, i, |t|) \rangle, g_{1,2} = \langle +(i,1,i) \rangle.$$

For a connected graph, its connection ($s_{2,2}$) is performed with successive symbols of the text-line, matching the top of the graph with the load t_i to the symbol of the text, $s_{2,1}$ ensures progress along the text-line.

$$s_{1,1} = s_{2,1} = s_{3,1},\ s_{2,2} = \langle \beta\ _{\tau 2} \rightarrow \Delta(r, t_i, G_f, v_s)\gamma \rangle,$$

$$g_{2,1} = \langle <(\tau 2, i, |t|) \rangle,$$

$$g_{2,2} = \langle +(i,1,i), =(s,i), =(r2,r) \rangle.$$

As long as there is a vertex in the graph for each subsequent character of the text, the constructor moves to the next graph vertex and the text character

$$s_{3,2} = \langle \gamma\ _{\tau 3} \rightarrow \Delta(r, t_i, G_f, v_s)\gamma \rangle,$$

$$g_{3,1} = \langle <(c, i, |t|), ==(r2, r2, true), \&(\tau 3, c, r2) \rangle,$$

$$g_{3,2} = \langle +(i,1,i), =(r2,r) \rangle.$$

If no match was found for the current character of the text-line among the graph vertices and before that several characters were processed, according to the rule $s_{4,2}$ constructor writes the processed fragment boundaries s to the list f, and then it goes to the next word in the text ($s_{4,1}$).

$$s_{4,1} = s_{5,1} = \langle \alpha\ _{\tau 4} \rightarrow t_i \alpha \rangle,$$

$$s_{4,2} = \langle \gamma\ _{\tau 5} \rightarrow \times(i, f, s)\sigma \rangle,$$

doi: https://doi.org/10.15802/stp2023/288958

$$g_{4,1} = \left\langle \begin{matrix} == (\tau 4, t_i, nbsp), == (r, r, \ false), \\ > (r1, s, i), \&(\tau 5, r, r1) \end{matrix} \right\rangle$$

$$g_{4,2} = \langle +(i, 1, i) \rangle .$$

If no match was found for the current string text character among the vertices of the graph and the previous symbols were not processed, we move to the next word from the text ($s_{5,1}$) and then link the text to another graph

$$s_{5,2} = \langle \beta_{\ \tau 6} \to \sigma \rangle$$

$$g_{5,1} = \langle == (r, r, \ false), == (r1, s, i), \&(\tau 6, r, r1) \rangle$$

$$g_{5,2} = \langle +(i, 1, i) \rangle .$$

Rule s_6 allows you to calculate the percentage of borrowings in the text

$$s_{6,1} = \varepsilon$$

$$s_{6,2} = \left\langle \begin{matrix} \alpha_{\ \tau 7} \to /(\downarrow (q_1, t, \ f, \ param), \\ (\downarrow (q_2, t, \left[0, (|t| - 1) \right], param) \end{matrix} \right\rangle$$

$$g_{6,1} = \langle == (\tau 7, i, |t|) \rangle .$$

The interpretation means the established correspondence between the operations of Σ_K and the algorithms of some algorithmic structure containing the set of algorithms $V_A = \left\{ A_i \middle| \begin{matrix} Y_i \\ X_i \end{matrix} \right\}$, where X_i, Y_i – are the sets of input and output values of the algorithm A_i.

The algorithm of the operation $\downarrow (q, t, \ f, \ param)$ consists of:

1. $q = 0, k = 0$;
2. until the end of the list f (k is less than the number of elements in f):
   2.1. for the text fragment t from the symbol f_k to f_{k+1} count the words number q_c;
   2.2. if $q_c \geq param$, then $q = q + q_c$;
   2.3. $q_c = 0, k = k + 2$,.

The implementation of the constructor means the formation of a set of values $\Omega(C_C) \in [0;1]$, which is the degree of borrowing of line texts concerning the texts in the graphic representation.

Experimental studies. Predefined constructive and production models of the texts authorship determining processes and their software implementations are applied to experimentally test the hypothesis regarding the possible statistical relationship between the results of solving the corresponding problems: the task of identifying borrowings and the task of text authorship establishing according to the style and other features of the author's text.

The purpose of the experiment. To determine the suitability of using the text authorship determining method with the help of a constructor that displays the sentence structure for the tasks of detecting borrowings (in a broader sense – plagiarism).

The experimental base is 16 text files in docx format, which are documentation for diploma projects of the OKR «Bachelor» in the direction 6.050103 «Software engineering» DNUZT-2018 (size 0.7 Mb – 27.3 Mb). Each file contains structural sections (28-33 pieces). Each section is allocated in a separate txt file. The total texts sections number (files) is 509.

The technical characteristics of the PC do not affect the results of the experiment.

Methodology of the experiment. The experiment consists of three logical parts:

1) determination of the borrowings percentage in the text using the graphical text representation model [12, 25];

2) determining the percentage of borrowings by analyzing the author's style;

3) results 1 and 2 correlation coefficient calculation.

Part I has the following stages:

1) automated analysis of the document structure, which is performed based on the analysis of the XML structure of its file, according to which the headings, designed with the help of built-in heading styles, determine the boundaries of sections [16]; formation of txt files containing the texts of individual sections. When creating files, the texts of the section undergo preliminary processing: removal of control symbols, conversion to one case, unification of punctuation marks, etc.;

2) the i-th document files-texts section graphic representation construction;

doi: https://doi.org/10.15802/stp2023/288958

3) setting the parameters and comparing the j-th document txt files set with the i-th document sections graphical representation;

4) forming a summary results table (table 1);

5) assignment of new comparison parameters values and points 3–4 repetition;

6) transition to the $(j+1)$ th document.

Part II consists of the following stages (add the constructor):

1) completely coincides with the first part of the experiment step 1;

2) conversion of the text from the txt-file into a formatted text with the parts-of-speech indication, number and gender using the first constructor-converter CP;

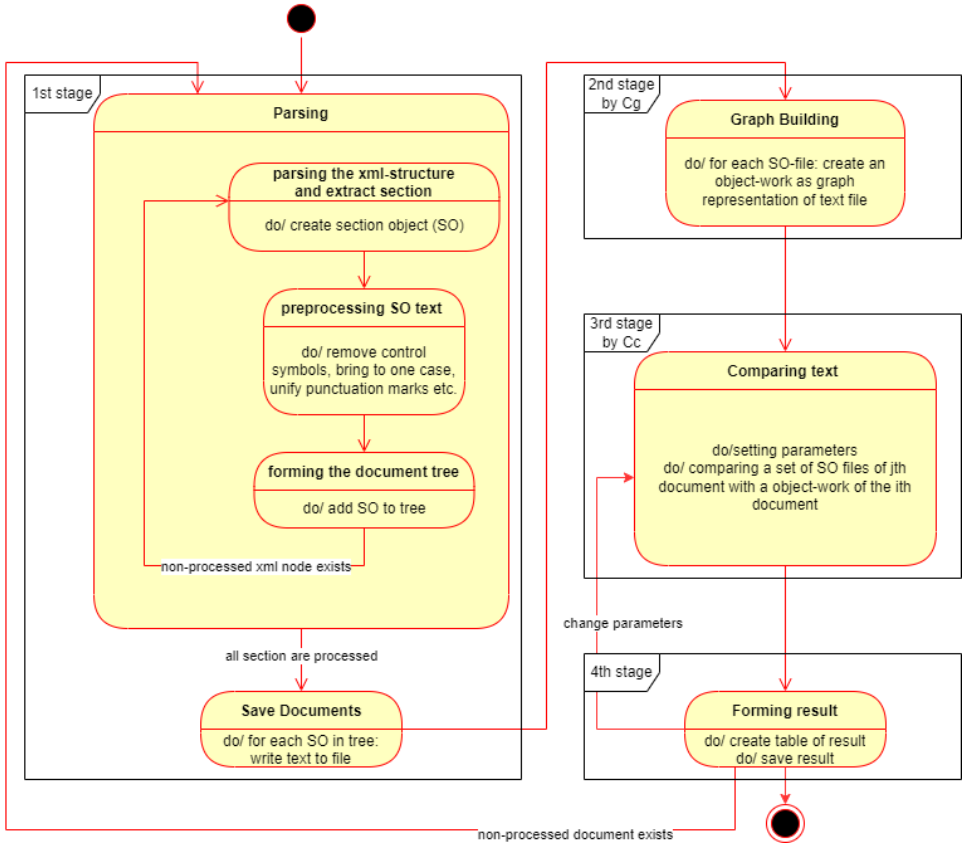

Fig. 1. The sequence of performing the search for borrowings in natural language texts using the graph constructor

3) forming the rules of the stochastic constructor based on the tagged text by the second constructor-converter CT;

4) with the help of a constructor-measurement CE, the calculation of the similarity of two stochastic constructors reflecting the syntactic structure of the texts being compared;

5) formation of a summary table of results;

Table 1 presents the sequential comparison results of the two diplomas' relevant sections with each other (P1, P2,..P30) using the two methods described above in a percentage of coincidence between them.

Due to the differences between the two approaches to comparison – the use of sentences in the first and words sequences without taking into account sentences in the second – the graph constructor worked with different comparison parameters, which are the type of fragment and its minimum length, at which a fragment can be considered borrowed (3–7 respectively).

The work result of the constructor-calculator based on the sentences' syntactic structure in the two relevant sections is located in the «sentence» row of the table. 1 and reduced to percentage form.

doi: https://doi.org/10.15802/stp2023/288958

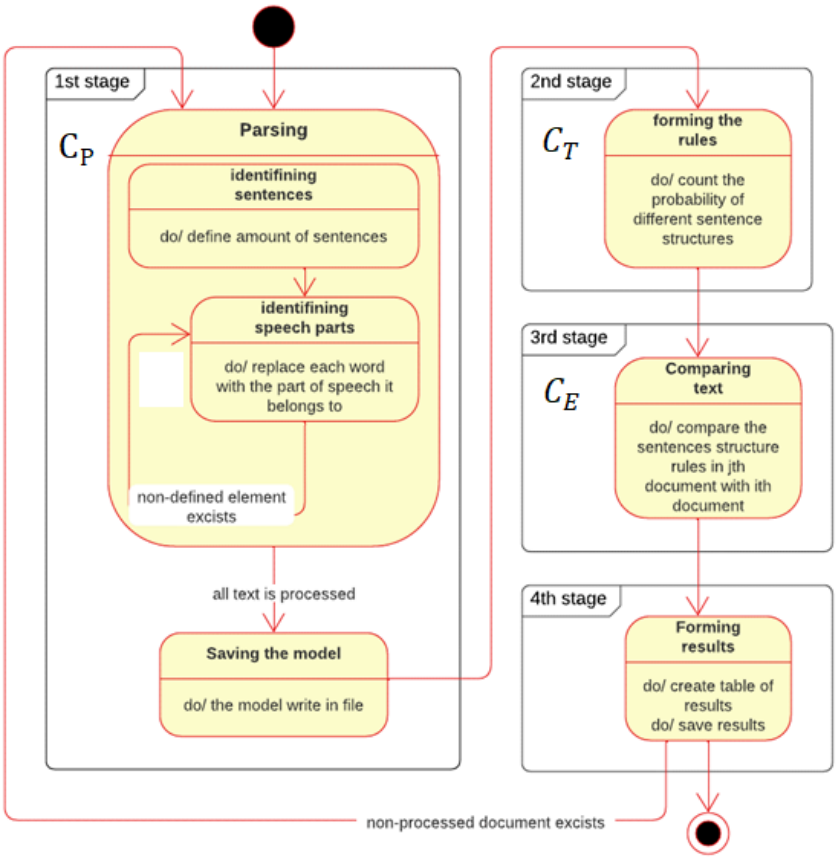

Fig. 2. The sequence of performing the search for borrowings in natural language texts using the sentence constructor

The section comparison result based on the graphical representation is located in lines 3–7 of the words in the table. 1, depending on the selected word sequence length for comparison. For clarity, the approximate percentage of section similarity is highlighted in the table.

Table 1

File comparison results using the graph constructor and the sentence structure constructor

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
sentence	**0.00**	**23.00**	**45.00**	3.00	17.00	**62.00**	**0.00**	**42.00**	**6.00**	2.00
3 words	**0.66**	27.43	62.79	5.56	21.43	65.04	**0.00**	**45.45**	**5.48**	6.67
4 words	**0.00**	**25.66**	55.81	0.00	0.00	**63.69**	**0.00**	**45.45**	**5.48**	6.67
5 words	**0.00**	**25.66**	**46.51**	0.00	0.00	**63.18**	**0.00**	**45.45**	**6.85**	0.00
6 words	**0.00**	**22.71**	**46.51**	0.00	0.00	**62.54**	**0.00**	0.00	0.00	0.00
7 words	**0.00**	**20.94**	**46.51**	0.00	0.00	**62.54**	**0.00**	0.00	0.00	0.00

doi: https://doi.org/10.15802/stp2023/288958

Continuation of Table 1

	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
sentence	**0.00**	60.00	**3.00**	**16.00**	**4.00**	**25.00**	**5.00**	**100.00**	**39.00**	2.00
3 words	24	35.73	**2.07**	**15.76**	9.69	41.38	4.51	**100.00**	**38.71**	29.63
4 words	**0.00**	68.43	**3.45**	**15.76**	**5.2**	0.00	4.51	**100.00**	**38.71**	29.63
5 words	**0.00**	31.86	**3.45**	13.79	**0.00**	**24.14**	4.51	**100.00**	**38.71**	0.00
6 words	**0.00**	30.47	0.00	13.79	**4.62**	15.52	4.51	**100.00**	**38.71**	0.00
7 words	**0.00**	28.81	0.00	13.79	**3.93**	15.52	0.00	**100.00**	0.00	0.00
	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30
sentence	**1.00**	**0.00**	**19.00**	**68.00**	**0.00**	**0.00**	**56.00**	**17.00**	**14.00**	**0.00**
3 words	**0.42**	2.00	20.95	22.92	**0.00**	**0.00**	60.37	**15.15**	24.24	16.67
4 words	**0.76**	3.33	**18.1**	19.79	**0.00**	**0.00**	**58.74**	**15.15**	**15.15**	**0.00**
5 words	0.00	3.33	**18.1**	**73.95**	**0.00**	**0.00**	**55.7**	**15.15**	**15.15**	**0.00**
6 words	0.00	**0.00**	**18.1**	19.79	**0.00**	**0.00**	52.99	0.00	0.00	**0.00**
7 words	0.00	**0.00**	**18.1**	19.79	10.67	**0.00**	47.77	0.00	0.00	**0.00**

Obtaining a zero similarity of some partitions is usually due to their size being too small to reliably reflect the similarity. In this work: fragment type – word, minimum length: from 3 to 7 words. This length is determined by the results of research [16], the data of which are partially shown in Table 2.

Since the experimental base of this study contains scientific style texts, which mainly have a complete grammatical basis and secondary clauses, the minimum sentence length is three words.

Regarding the maximum length, based on the data in the table. 2, it is advisable to take 7–9. However, the parallel execution of the experiment's second part indicates the consideration sufficiency of the maximum equal to seven.

Discussion of experimental results

The obtained similarity results for 16 diploma theses were compared and the correlation coefficient was obtained for the results of the work of the two described approaches. The comparison was made taking into account the different lengths of the sequence of words, from 3 to 7 in number, and the following results were obtained.

For 3 words in a row, the average value of the correlation coefficient was 0.00053, which is an unsatisfactory result and demonstrates a large discrepancy between the results of the two applied methods.

When using a sequence of words with a length of 4, the reliability of the results has improved significantly – the average value is 0.82, which allows us to say that the analysis starting with 4 words is reliable and reflects the real state of affairs.

In longer experiments using word lengths 5 and 6, the obtained results also reflect the feasibility of using precisely these lengths of word sequences as the most informative. The average value of the correlation coefficients is 0.88 for calculations with a sequence length of 5 and 0.82 for a length of 6 words.

The results of working with a 7 words sequence. The result is similar to working with 3 words in a row and it points out the impracticality of their use

The average value of the correlation coefficient is 0.000531, which is an unsatisfactory result and indicates a strong discrepancy between the two methods. The general result can be considered the sufficient correlational similarity of the two methods and identification of the required sequence lengths for a reliable reflection of the author's style.

doi: https://doi.org/10.15802/stp2023/288958

Table 2

Lengths of sentences that are used most often

Length in words	Drama, %	Artistic prose, %	Poetry, %
1–3	49.73		
4–6	29.07	18.6	18.87
7–9	12.14	18.65	23.02
10–12		18.33	18.33

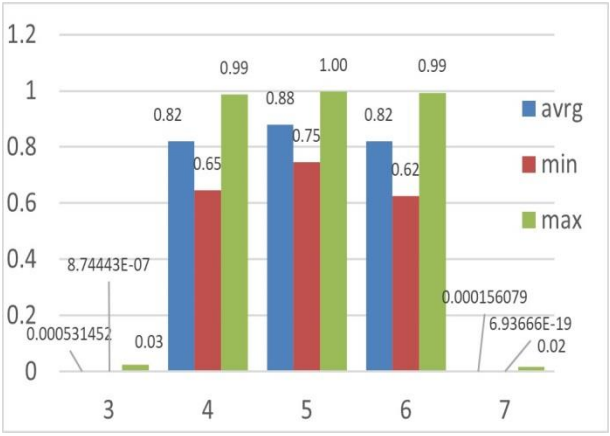

Fig. 3. Maximum and minimum correlation similarity values of two methods for 16 theses using sequences of length 3–7

Table 3

Correlational similarity of the results obtained using the constructor of the sentence structure and the constructor of graphs with 5 words in a row for 16 diploma theses

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		96	91	84	77	84	89	89	97	81	81	83	94	85	97	82
2	98		95	88	86	96	80	90	86	86	100	75	77	85	79	83
3	97	99		82	81	94	98	79	88	86	76	83	91	79	96	83
4	90	98	80		89	75	86	77	76	80	97	84	98	94	87	99
5	88	89	95	95		82	81	89	99	94	91	86	85	75	78	87
6	95	89	85	82	86		99	91	98	83	97	89	96	77	98	88
7	98	84	91	84	99	82		89	87	87	99	78	86	91	81	79
8	88	93	99	75	84	83	91		81	93	89	76	89	89	80	98
9	90	89	92	86	87	91	99	86		84	87	97	75	96	88	87
10	97	99	87	86	85	75	87	98	100		98	93	75	98	79	83
11	91	93	79	87	92	90	90	76	86	89		95	84	77	93	82
12	90	88	91	89	83	91	98	86	76	93	75		91	89	94	84
13	88	99	78	80	85	93	95	83	91	96	95	98		94	84	78
14	94	99	87	91	95	80	90	79	98	81	86	82	97		98	75
15	96	87	94	86	94	78	98	79	99	87	96	88	85	82		98
16	99	95	76	93	79	89	90	76	90	82	96	76	91	75	78	

doi: https://doi.org/10.15802/stp2023/288958

Originality and practical value

The research was carried out on technical programming texts. It is expected that the method can also be applied to texts from other technical fields, but this position should be supported by relevant experiments.

With small volumes of the text, a weak results correlation for identifying borrowings and establishing the texts authorship was observed, which requires further research. There are no clear boundaries between small and large volumes of text. According to the results of the experiments, it was established that for a satisfactory result, the minimum volume of text should be 10 characters and consist of 5 sentences.

When choosing text samples, the author should take into account that the author's style can change due to the passage of time, different text topics, and changes in commonly used templates for the text's formation.

We believe that the proposed method of the text's authorship establishing can be widespread and effectively used for various Slavic languages. Other languages, such as English, where fewer attributes of words, which are compensated by sentence building patterns, as well as more formal requirements for it, significantly weaken the capabilities of the proposed methods.

The presented method of the text's authorship establishing can also be used to identify the presence of a large borrowings volume. This can serve as a reason for further, more thorough verification by software tools or with the involvement of experts.

To check the text for the presence of plagiarism, it is necessary to have a large bank of texts by other authors to detect these borrowings, which can be difficult due to the constant increase in the number of materials in free access and the variety of forms and formats of their presentation. Unlike well-known programs for identifying borrowings and establishing authorship, the proposed approach is limited to the availability of a relatively small volume of the author's texts and does not require a large bank of texts.

In the course of working with explanatory notes for students' diplomas in the programming field, it was established that the approach allows working only with natural language text. The method did not work with sections that include program code that could not be processed this way. In the future, it is planned to develop a constructor that will be able to process texts in formal languages.

Findings

The work confirms the hypothesis regarding the high connection between tasks, methods of solving them, and results regarding establishing the authorship of technical texts and identifying borrowings. It was established that the correlation ratio between the results can be more than 0.9%.

A constructive-production model of the text authorship establishment was developed based on the features analysis and regularities of the author's sentence formation style. The essence of the model is to formalize the representation of the author's sentence syntax by a set of substitution rules with a probability load. The obtained results show that the proposed method has high efficiency compared to the methods used earlier [3].

A model of technical texts was developed taking into account the author's style, thanks to the reflection of the unique stylistic and linguistic features of the author's own language allows to significantly simplify the identifying borrowings process and establishing authorship due to the using only one author's work instead of a whole texts' corpus – possible sources of borrowing.

The results of the experiments determined the value of the rational parameter – the text fragments minimum length (in the number of words) that should be considered borrowing, it is equal to five words. This should be considered as a recommendation for the use of any plagiarism detection programs.

The proposed method can be used both to solve the problems of finding borrowings and to establish the probable authorship of the text.

doi: https://doi.org/10.15802/stp2023/288958

LIST OF REFERENCE LINKS

1. Кульчицький І. М. Дослідження довжини речення та слова у творах Романа Іваничука. *Вісник Національного університету Львівська політехніка. Серія : Інформаційні системи та мережі*. 2017. № 872. С. 139–148.
2. Плющ М. Я. *Граматика української мови. Морфеміка. Словотвір. Морфологія*. Київ : Видавничий дім «Слово», 2010. 328 с.
3. Шинкаренко В. И., Куропятник Е. С. Конструктивно-продукционная модель графового представления текста. *Теоретичні та методологічні основи програмування.* 2016. № 2–3. С. 63–72. DOI: https://doi.org/10.15407/pp2016.02-03.063
4. Ahuja L., Gupta V., Kumar R. A New Hybrid Technique for Detection of Plagiarism from Text Documents. *Arabian Journal for Science and Engineering.* 2020. Vol. 45. Iss. 12. P. 9939–9952. DOI: https://doi.org/10.1007/s13369-020-04565-9
5. Al-Smadi M., Jaradat Z., Al-Ayyoub M., Jararweh Y. Paraphrase identication and semantic text similarity analysis in arabic news tweets using lexical, syntactic, and semantic features. *Information Processing & Management.* 2017. Vol. 53. Iss. 3. P. 640–652. DOI: https://doi.org/10.1016/j.ipm.2017.01.002
6. Ceska Z. Plagiarism detection based on singular value decomposition. *In Advances in Natural LanguageProcessing*. 2008. P. 108–119. DOI: https://doi.org/10.1007/978-3-540-85287-2_11
7. Demidovich I., Shynkarenko V., Kuropiatnyk O., Kirichenko O. Processing Words Effectiveness Analysis in Solving the Natural Language Texts Authorship Determination Task. *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)* (Lviv, 22-25 Sept. 2021). Lviv, 2021. P. 48–51. DOI: https://doi.org/10.1109/CSIT52700.2021.9648829
8. Eyecioglu A., Keller B. Twitter paraphrase identication with simple overlap features and SVMs. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 2015. P. 64–69. DOI: https://doi.org/10.18653/v1/s15-2011
9. Foltýnek T., Meuschke N., Gipp B. Academic Plagiarism Detection: A Systematic Literature Review. *ACM Computing Surveys.* 2019. Vol. 52. Iss. 6. P. 1–42. DOI: https://doi.org/10.1145/3345317
10. Gillam L., Vartapetiance A. From English to Persian: Conversion of Text Alignment for Plagiarism Detection. *FIRE (Working Notes).* 2016. P. 160–162.
11. Gómez-Adorno H., Sidorov G., Pinto D., Markov I. A graph based authorship identification approach. *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15).* 2015. P. 1–7.
12. Güllü M., Polat H. Text Authorship Identification Based on Ensemble Learning and Genetic Algorithm Combination in Turkish Text. *Politeknik Dergisi.* Vol. 25. Iss. 3. P. 1287–1297. DOI: https://doi.org/10.2339/politeknik.992493
13. Gupta D., Vani K., Singh C. K. Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection. *International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (Delhi, 24-27 September 2014). Delhi, 2014. P. 2694–2699. DOI: https://doi.org/10.1109/icacci.2014.6968314
14. Hussain F., Suryani M. On retrieving intelligently plagiarized documents using semantic sim-ilarity. *Engineering Applications of Artificial Intelligence.* 2015. Vol. 45. P. 246–258. DOI: https://doi.org/10.1016/j.engappai.2015.07.011
15. Kuropiatnyk O., Shynkarenko V. Automation of template formation to identify the structure of natural language documents. *COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems* (Kharkiv, April 22–23 2021). Kharkiv, 2021. P. 179–190.
16. Lupei M., Mitsa A., Repariuk V., Sharkan V. Identification of Authorship of Ukrainian-Language Texts of Journalistic Style Using Neural Networks. *Eastern-European Journal of Enterprise Technologies.* 2020. Vol. 1. Iss. 2 (103). P. 30–36. DOI: https://doi.org/10.15587/1729-4061.2020.195041
17. Meuschke N., Gondek Ch., Seebacher D., Breitinger C., Keim D. A., Gipp B. An adaptive image-based plagiarism detection approach. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries.* 2018. P. 131–140. DOI: https://doi.org/10.1145/3197026.3197042
18. Meuschke N., Schubotz M., Hamborg F., Skopal T., Gipp B. Analyzing mathematica lcontent to detect academic plagiarism. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* 2017. P. 2211–2214. DOI: https://doi.org/10.1145/3132847.3133144

doi: https://doi.org/10.15802/stp2023/288958

19. Meuschke N., Stange V., Schubotz M., Kramer M., Gipp B. Improving academic plagiarism detection for STEM documents by analyzing mathematical content and citations. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL).* 2019. P. 120–129. DOI: https://doi.org/10.1109/JCDL.2019.00026
20. Najafi M., Ehsan T. Text-to-Text Transformer in Authorship Verification Via Stylistic and Semantical Analysis. *CLEF 2022 – Conference and Labs of the Evaluation Forum* (Bologna, 5–8 September 2021). Bologna, 2021. P. 1–10.
21. Rakian S., Safi E. F., Rastegari, H. A Persian fuzzy plagiarism detection approach. *Journal of Information Systems and Telecommunication.* 2015. Vol. 3, No. 3. P. 182–190. DOI: https://doi.org/10.7508/jist.2015.03.007
22. Satyapanich T., Gao H., Finin T. Ebiquity: Paraphrase and semantic similarity in twitter usingskipgrams. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).* 2015. P. 51–55. DOI: https://doi.org/10.18653/v1/s15-2009
23. Shynkarenko V. I., Demidovich I. M. Constructive-synthesizing modeling of natural language texts. *Computer systems and information technologies.* 2023. № 3. (Preprint).
24. Shynkarenko V. I., Demidovich I. M. Natural Language Texts Authorship Establishing Based on the Sentences Structure. *COLINS-2022 : 6th International Conference on Computational Linguistics and Intelligent Systems* (Gliwice, 12–13 May 2022). Gliwice, 2022. P. 328–337.
25. Shynkarenko V. I., Ilman V. M. Constructive-Synthesizing Structures and Their Grammatical Interpretations. i. Generalized Formal Constructive-Synthesizing Structure. *Cybernetics and Systems Analysis.* 2014. Vol. 50. Iss. 5. P. 655–662. DOI: https://doi.org/10.1007/s10559-014-9655-z
26. Shynkarenko V., Kuropiatnyk O. Constructive Model of the Natural Language. *Acta Cybernetica.* 2018. Vol. 23. Iss. 4. P. 995–1015. DOI: https://doi.org/10.14232/actacyb.23.4.2018.2
27. Tschuggnall M., Specht G. Using Grammar-Profiles to Intrinsically Expose Plagiarism in Text Documents. *Natural Language Processing and Information Systems.* 2013. Vol. 7934. P. 297–302. DOI: https://doi.org/10.1007/978-3-642-38824-8_28

В. І. ШИНКАРЕНКО[1*], І. М. ДЕМИДОВИЧ[2*], О. С. КУРОП'ЯТНИК [3*]

[1*]Каф. «Комп'ютерні інформаційні технології», Український державний університет науки і технологій, вул. Лазаряна, 2, Дніпро, Україна, 49010, тел. +38 (056) 373 15 52, ел. пошта shinkarenko_vi@ua.fm, ORCID 0000-0001-8738-7225
[2*]Каф. «Комп'ютерні інформаційні технології», Український державний університет науки і технологій, вул. Лазаряна, 2, Дніпро, Україна, 49010, тел. +38 (056) 373 15 52, ел. пошта 2019demidovichinn@gmail.com, ORCID 0000-0002-3644-184X
[3*]Каф. «Комп'ютерні інформаційні технології», Український державний університет науки і технологій, вул. Лазаряна, 2, Дніпро, Україна, 49010, тел. +38 (056) 373 15 52, ел. пошта olena.kuropiatnyk@gmail.com, ORCID 0000-0003-2286-884X

Подвійний підхід до встановлення авторства технічних природномовних текстів та їх складових

Мета. Дослідження спрямовано на перевірку гіпотези щодо можливостей визначення плагіату методами встановлення авторству текста без використання банку текстів та їх безпосереднього порівняння. **Методика.** Розроблено конструктивно-продукційні моделі процесів встановлення авторства технічних текстів для двох методів. Перший метод заснований на формуванні моделі тексту у вигляді безлічі формальних правил підстановки з імовірнісними вагами (як у стохастичних формальних граматиках), що відображає синтаксичні особливості та закономірності формування тексту автором. Встановлюється ступінь схожості досліджуваного тексту з іншим методом порівняння їх моделей. Другий метод – класичний підхід до виявлення запозичень (плагіату) шляхом безпосереднього порівняння досліджуваного тексту з наявним банком текстів, виділення фрагментів тексту, що повторюються, і встановлення ступеня оригінальності. Виконано експерименти щодо встановлення кореляційної залежності результатів цих двох методів. Експериментальна база складалася з 509 текстових секцій дипломних робіт студентів спеціальності «Програмна інженерія». **Результати**. Експериментальні дослідження дали змогу встановити високу кореляційну залежність між результатами двох методів. Коефіцієнт кореляції в межах 0,75…1,0 та із середнім значенням 0,88 отримано за умови, що запозичення враховуються для фрагментів тексту завдовжки не менше п'яти слів. **Наукова новизна.** Автори вперше встановили можливості та запропонували методи опосередкованого виявлення плагіату без використання банку текстів значного обсягу. Суть моделі полягає у формалізації

doi: https://doi.org/10.15802/stp2023/288958

представлення синтаксису речення автора набором правил підстановки з імовірнісними вагами. **Практична значимість**. На основі отриманих результатів розширено можливості з виявлення запозичень та підвищено результативність відповідних методів. Отримано рекомендації щодо параметрів класичних методів виявлення запозичень, зокрема як раціональний параметр під час використання систем виявлення запозичень рекомендовано враховувати фрагменти тексту довжиною не менше п'яти слів. Поширено можливості методів встановлення авторства текстів, апробованих на текстах художньої літератури, на використання і для технічних текстів.

Ключові слова: природномовні тексти; визначення авторства; статистичний аналіз; класифікація; коефіцієнт кореляції; конструктивно-продукційне моделювання; конструктивізм; формальні граматики; графи

REFERENCES

1. Kulchytskyi, I. M. (2017). The examination of sentence and word length in the writing of Roman Ivanychuk. *Bulletin of Lviv Polytechnic National University series: «Information Systems and Networks»*, 139-148. (in Ukrainian)
2. Pliushch, M. Ya. (2010). *Hramatyka ukrainskoi movy. Morfemika. Slovotvir. Morfolohiia*. Kyiv: Vydavnychyi dim «Slovo». (in Ukrainian)
3. Shynkarenko, V. I., & Kuropiatnyk, O. S. (2016). Constructive-synthesizing model of text graph representation. *Problems in programming, 2-3*, 63-72. DOI: https://doi.org/10.15407/pp2016.02-03.063 (in Russian)
4. Ahuja, L., Gupta, V., & Kumar, R. (2020). A New Hybrid Technique for Detection of Plagiarism from Text Documents. *Arabian Journal for Science and Engineering, 45*(12), 9939-9952. DOI:https://doi.org/10.1007/s13369-020-04565-9 (in English)
5. AL-Smadi, M., Jaradat, Z., AL-Ayyoub, M., & Jararweh, Y. (2017). Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features. *Information Processing & Management, 53*(3), 640-652. DOI: https://doi.org/10.1016/j.ipm.2017.01.002 (in English)
6. Ceska, Z. (2008). Plagiarism Detection Based on Singular Value Decomposition. *Lecture Notes in Computer Science*, 108-119. DOI: https://doi.org/10.1007/978-3-540-85287-2_11 (in English)
7. Demidovich, I., Shynkarenko, V., Kuropiatnyk, O., & Kirichenko, O. (2021, September). Processing Words Effectiveness Analysis in Solving the Natural Language Texts Authorship Determination Task. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)* (pp. 48-51). Lviv, Ukraine. DOI: https://doi.org/10.1109/csit52700.2021.9648829 (in English)
8. Eyecioglu, A., & Keller, B. (2015). Twitter Paraphrase Identification with Simple Overlap Features and SVMs. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 64-69. DOI: https://doi.org/10.18653/v1/s15-2011 (in English)
9. Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic Plagiarism Detection. *ACM Computing Surveys, 52*(6), 1-42. DOI: https://doi.org/10.1145/3345317 (in English)
10. Gillam, L., & Vartapetiance, A. (2016). From English to Persian: Conversion of Text Alignment for Plagiarism Detection. *FIRE (Working Notes)*, 160-162. (in English)
11. Gómez-Adorno, H., Sidorov, G., Pinto, D., & Markov, I. (2015). A graph based authorship identification approach. *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15),* 1-7. (in English)
12. Güllü, M., & Polat, H. (2022). Text Authorship Identification Based On Ensemble Learning and Genetic Algorithm Combination in Turkish Text. *Politeknik Dergisi, 25*(3), 1287-1297. DOI: https://doi.org/10.2339/politeknik.992493 (in English)
13. Gupta, D., Vani, K., & Singh, C. K. (2014). Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2694-2699). Delhi, India. DOI: https://doi.org/10.1109/icacci.2014.6968314 (in English)
14. Hussain, S. F., & Suryani, A. (2015). On retrieving intelligently plagiarized documents using semantic similarity. *Engineering Applications of Artificial Intelligence, 45*, 246-258. DOI: https://doi.org/10.1016/j.engappai.2015.07.011 (in English)
15. Kuropiatnyk, O., & Shynkarenko, V. (2021, April). Automation of template formation to identify the structure of natural language documents. In *COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems* (pp. 179-190). (in English)

doi: https://doi.org/10.15802/stp2023/288958

16. Lupei, M., Mitsa, A., Repariuk, V., & Sharkan, V. (2020). Identification of authorship of Ukrainian-language texts of journalistic style using neural networks. *Eastern-European Journal of Enterprise Technologies, 1*(2(103)), 30-36. DOI: https://doi.org/10.15587/1729-4061.2020.195041 (in English)
17. Meuschke, N., Gondek, C., Seebacher, D., Breitinger, C., Keim, D., & Gipp, B. (2018). An Adaptive Image-based Plagiarism Detection Approach. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 131-140). DOI: https://doi.org/10.1145/3197026.3197042 (in English)
18. Meuschke, N., Schubotz, M., Hamborg, F., Skopal, T., & Gipp, B. (2017). Analyzing Mathematical Content to Detect Academic Plagiarism. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2211-2214). DOI: https://doi.org/10.1145/3132847.3133144 (in English)
19. Meuschke, N., Stange, V., Schubotz, M., Kramer, M., & Gipp, B. (2019). Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 120-129). DOI: https://doi.org/10.1109/JCDL.2019.00026 (in English)
20. Najafi M., Ehsan T. (2021, September). Text-to-Text Transformer in Authorship Verification Via Stylistic and Semantical Analysis. In *CLEF 2022-Conference and Labs of the Evaluation Forum* (pp. 1-10). Bologna, Italy. (in English)
21. Rakian S., Safi E. F., & Rastegari, H. (2015). A Persian fuzzy plagiarism detection approach. *Journal of Information Systems and Telecommunication, 3*(3), 182-190. DOI: https://doi.org/10.7508/jist.2015.03.007 (in English)
22. Satyapanich, T., Gao, H., & Finin, T. (2015). Ebiquity: Paraphrase and Semantic Similarity in Twitter using Skipgrams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 51-55. Gliwice, Poland. DOI: https://doi.org/10.18653/v1/s15-2009 (in English)
23. Shynkarenko, V. I., & Demidovich, I. M. (2023). Constructive-synthesizing modeling of natural language texts. *Computer systems and information technologies, 3.* (Preprint) (in English)
24. Shynkarenko, V. I., & Demidovich, I. M. (2022, May). Natural Language Texts Authorship Establishing Based on the Sentences Structure. In *COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems* (pp. 328-337). Gliwice, Poland. (in English)
25. Shynkarenko, V. I., & Ilman, V. M. (2014). Constructive-Synthesizing Structures and Their Grammatical Interpretations. i. Generalized Formal Constructive-Synthesizing Structure. *Cybernetics and Systems Analysis, 50*(5), 655-662. DOI: https://doi.org/10.1007/s10559-014-9655-z (in English)
26. Shynkarenko, V., & Kuropiatnyk, O. (2018). Constructive Model of the Natural Language. *Acta Cybernetica, 23*(4), 995-1015. DOI: https://doi.org/10.14232/actacyb.23.4.2018.2 (in English)
27. Tschuggnall, M., & Specht, G. (2013). Using Grammar-Profiles to Intrinsically Expose Plagiarism in Text Documents. *Natural Language Processing and Information System, 7934*, 297-302. DOI: https://doi.org/10.1007/978-3-642-38824-8_28 (in English)

Надійшла до редколегії: 02.02.2023
Прийнята до друку: 05.06.2023

doi: https://doi.org/10.15802/stp2023/288958