Kamil R.Ayda-zade[1], Sakhavat G.Talibov[2] DOI: 10.25045/jpit.v08.i1.02
Institute of Control Systems of ANAS, Baku, Azerbaijan
[1]Kamil_ayda-zade@rambler.ru, [2]saxavat@yahoo.com

ANALYSIS OF THE METHODS FOR THE AUTHORSHIP IDENTIFICATION OF THE TEXT IN THE AZERBAIJANI LANGUAGE

The methods and algorithms used for recognition texts authorship analyzes in the paper. The applied features of recognition are based on n-grams with n = 1, and n = 2. The results of computer experiments to identify the authorship of the texts in the Azerbaijani are presented.

Keywords: *identification, authorship identification, recognition, n-gram, support vector machine.*

1. Introduction

As it is known, one of the important problems of text mining is its classification by the authors, i.e., determining the alleged author of the particular text out of a predetermined group of authors.

In the 70s of the last century, researchers began to intensively focus on the automated solutions of this problem emerged. Initially, the methods to solve this problem based on the use of special glossaries created for the keywords.

Mosteller was one of the first using the Bayesian analysis to solve the authorship recognition problem [1]. Later, Burrows used the frequencies of the most frequently used words by the authors in his work [2], while Brinegar [3] employed the length of the words, Morton [4] - the length of sentences, Brainerd [5] - the average number of syllables, Holmes [6] - the number of the used words and the volume of the document, Twedie [7] - the ratio of the number of the used words to the total number of words in the text, Fürnkranz [9] and Tan [10] - n-gram (2-gram or 3-gram). Çatal [11] created the identification system NECL. In [18, 19], the frequency of occurrence of letters and letter combinations was used to recognize the authors of the Russian literature.

In [12] and [13], the frequency of the use of letters and the length of the words was first studied for the recognition of the authorship of the Azerbaijani texts. Nevertheless, there is no a computer system for the authorship recognition of the Azerbaijani texts yet. In this paper, we study the problem of authorship identification based on the analysis of author articles of small size (volume). The main challenge of the text (articles) authorship recognition of small volume in the Azerbaijani language is that a large number of uninformative suffixes and endings are used in the words. And the automated analysis of the words into the component parts for the Azerbaijani language still have not been solved.

2. Statement of the Problem

Formally, the statement of the problem of the texts authorship identification can be described as follows.

The database contains some texts of n authors and each of them has m_i number of texts of $D_{i,j,}$ j = 1, ..., m_i, i = 1, ..., n. The class (group) of the text of the i-th author is denoted by Y_i.The problem reviewed in the article is to determine the author of a new entered text D among the n authors, in other words to which class Y_i it belongs to. Each of the texts of $D_{i,j,}$ and D is associated with a set of values attributes.

We introduce the following notations, definitions and formulas.

Let's compare each of the texts $\mathcal{D}_{i,j}$and D to the set of characteristic values $\{M_{i,j}^{s}, \mathrm{s} \in K_i\}$ and $\{d_s,\ \mathrm{s} \in K_i,\}, K = \bigcap_{s=1}^{n} K_s,\ i = 1, \dots, n, j = 1 \dots, m_i,$ on the basis of which the classification of text by the authoroccurs. Where, K_i–is the set of attributes foridentification of the authorship of the i-th author, $i = 1, \dots, n$.

Let $N_{i,j}$- be the length (volume) of the j-th text of the i-th author, m_i–the number of articles of the i-th author contained in the database. Then it is clear that the average value of the s-th attribute of the i-th author in the j-th article is defined by the following formula:

$$\varepsilon_{i,j}^{s} = \frac{M_{i,j}^{s}}{N_{i,j}}, \quad s \in K_i, j = 1, \dots, m_i, \ \ i = 1, \dots, n, \tag{1}$$

the average value of the *s*-th attribute in all the articles of the *i*-th author equals to

$$\xi_i^s = \frac{\sum_{j=1}^{m_i} M_{i,j}^{s}}{\sum_{j=1}^{m_i} N_{i,j}}, \ \ s \in K_i, i = 1, \dots, n. \tag{2}$$

Let the average value of the *s*-th attribute in the new article *D* equal to

$$x_D^s = \frac{m_D^s}{N_D}, \qquad s \in K_i. \tag{3}$$

Here, m_D^s- is the value of the *s*-th attribute in the new article *D*, and N_D- its length (volume).

Obviously, the dispersion of the *s*-th attribute for the *i*-th author equals to

$$(d_i^s)^2 = \frac{\sum_{j=1}^{m_i} (M_{i,j}^{s} - \xi_i^s)^2}{\sum_{j=1}^{m_i} N_{i,j}}, \ \ s \in K_i, \ \ i = 1, \dots, n. \tag{4}$$

The variation of the *s*-th attribute for the *i*-th author equals to

$$v_i^s = \frac{d_i^s * 100}{\xi_i^s}, \ \ s \in K_i, \ \ i = 1, \dots, n. \tag{5}$$

Let's review the length

$$R_i = \sum_{s \in K} \alpha_s (x_D^s - \xi_i^s)^2, i = 1, \dots, n, \tag{6}$$

which determines the proximity (norm) of the values of the attributes of the new text *D* with the values of the attributes that characterize the *i*-th author; α_s–the weight (importance) of the *s*-th attribute to determine the authorship of the articles.

3. Classification of authors' identification computer systems

The following types of authors' identification computer systems are known:

- *the systems, not taking into account the language of the texts;*
- *the system, taking into account the language of the texts;*
- *combined recognition system with a hierarchical structure.*

Available identification systems, which do not take into account the language of the texts, as a rule, use the following attributes: the length of sentences; frequency of the use of paragraphs; the length of words; the length of sub-headings; the average size of paragraphs; the number of punctuation signs; the frequency of the most common words used by the author; the suffixes most commonly used by the author and others.

Many existing computer identification systems that do not take into account the language of the texts, based on the use of methods that analyze the frequency of use of various combinations of letters consisting of the n-letters of n-grams.

The use of n-grams for many languages has its difficulties, especially for context-free languages and the languages in which the root of the word can be used with several suffixes, such as the Turkic (Turkish, Azeri, Uzbek, etc.), Slavic (Russian, Belarusian, Polish, Slovak, etc.) language groups. For example, the derivative words in the Azerbaijani language can be used with the prefix, suffix, and so on. The use of the most frequently used words to identify the authorship in computer systems requires the separation of the root from the suffix. Unfortunately, as mentioned above, this problem for the Azerbaijani language has not been solved yet. In addition, it is important to ignore articles, dividers, special words, figures, and the most frequently used words (stop words) to identify the authorship.

Identification systems, taking into account the text language, generally use the following features:

the frequency of the use of stop word in the text; the use of certain parts of speech (e.g., noun, pronoun, etc.); the frequency of commonly used prefixes and suffixes by the author, and others.

Combined systems of hierarchical identification structure are created on the basis of several recognition stages (levels). If the text is not artificial, the efficiency of using the allocation of suffixes for one language may be wrong for another language and provide a completely erroneous result. For example, in English or German, the prefix reduction less significantly affects the authorship recognition of the text, but it is important, particularly for the Azerbaijani and Turkish languages.

It should be noted that the used certain accents and dialects of the language are also the key factors influencing the identification of an author, although in some cases (e.g., artistic and scientific works), the use of accents and dialects does not really matter. The practice of the development of the texts authorship identification systems shows that the use of combination algorithms of hierarchical structure can significantly improve the efficiency of the text authorship recognition system.

4. Methods and algorithms used for text authorship recognition

This article describes the developed algorithms that do not use the specifics of language and provides the results of their performance in terms of identifying the authorship of the articles taken from the newspapers and news sites in the Azerbaijani language.

It should be noted that the major problem of obtaining high-quality results of the authorship recognition with these algorithms is the assumption of a small size of the available articles of each author. At the same time, a large part of the information contained in the available articles, as a rule, is related to the specific topics, i.e., specific words, terms, which are defined by a specific theme, are used. Therefore, it is not sufficiently informative for the authorship recognition.

The algorithms developed for the authorship identification methods generally include the implementation of the following series of processes:

- held initial processing of existing texts (articles, works) of different authors, and the numerical values of the selected attributes are defined each author;
- the attribute values are analyzed and the set of informative attributes is determined for each author (the set of attributes can differ for various authors);
- the attribute values presented to the new article by an unknown author are determined;
- the supposed author of the given article is defined according to specific criteria based on known algorithms.

Here are the signs of the authorship based on statistical analysis of the letter combinations. It should be noted that many available algorithms and text authorship recognition systems apply the attributes based on the analysis of the use of various letters combinations of n-letters, known as n-grams in the literature. Thus, in this case, gram means that one letter is taken as a unit, wherein the words, sentences or paragraphs of entire text are split into the letter combinations depending on the value n. These letter combinations comprise successive letters n, n = 1,2, ... For example, the phrase “Баку − столица” (“Baku - capital”) is split into monograms (n equals to 1 letter) and diagrams (n equals to 2 letters), respectively, as follows:

“Б”, “а”, “к”, “у”, “_”,“с”, “т”, “о”, “л”, “и”, “ц”, “а” ,
“Ба”, “ак”, “ку”, “у_”, “_с”, “ст”, “то”, “ол”, “ли”, “иц”, “ца”.
(“B”, “a”, “k”, “u”, “_”, “c”, “a”, “p”, “i”, “t”, “a”, “l”,
“Ba”, “ak”, “ku”, “u”,“_”, “c”, “ca”, “ap”, “pi”, “ta”, “al”.)
Here,“_”- denotes the space.

Noteworthy that the number of letters, and consequently, there are 32 1-grams in the Azerbaijani language, but a practical number of possible 2-grams equals to 835. The following three algorithms using the attributes based on *n*-grams have been realized.

Here is the general description of the algorithm 1, using monograms.

Step 1. The frequency of use of all the letters of the alphabet (1 gram) is determined by the formula (1) as an attribute for each text (article) of the i-th author included in the class Y_i.

Step 2: Combining all the articles of each author in accordance with the formula (2), the average values of all the attributes are calculated.

Step 3: A vector of the values of the attributes$x_{\mathcal{D}}^{s}$is calculated by the formula (3) for the new explored article D, s ∈ K.

Step 4. In the formula (6), taking $\alpha_s = 1$ (all the weights equal to 1), we define such ν, that $R_\nu = \min\limits_{1\le i\le n} R_i$, accordingly, the author of the article D is the ν-th author.

The diagram based algorithm 2 we used is the same as the algorithm 1 for monograms, but it does not analyze the frequency of the used individual letters, but the frequency of the used various combinations of two letters of the alphabet (2-gram - diagram), used by the author in his articles.

Here is the description of the modified algorithm 3, which uses the monograms. The basic idea of the algorithm, which is based on the modified monograms, is that, in this case, it uses sustainable attributes not including the sets of uncharacteristic attributes of the author.

The proposed algorithm calculates the variation of each attribute for each author using the formula (5). The weight of the s-th attribute in the formula (6) is defined depending on the variation value of the s-th attribute by all authors as follows:

$$v^s = \begin{cases} \min\limits_i v_i^s, & \min\limits_i v_i^s > \varepsilon, \\ \varepsilon, & \min\limits_i v_i^s \le \varepsilon, \end{cases} \quad s \in K.$$

Supposed length ε is defined basing on the length of the value of the attributes that are non-attributive for the authors. Then

$$\propto_s = \frac{(v^s)^{-1}}{\sum_{j=1}^{k}(v^j)^{-1}}, \quad s \in K.$$

It is clear that $\propto_s$ provides the following conditions:

$$0 \le \propto_s \le 1 \quad , \quad \sum_{s\in K} \propto_s = 1 \quad .$$

The first two steps of the proposed algorithm are same as the two steps of the first algorithm.

Step 3. Using (4) and (5), the resistance of the attributes is checked basing on the variation for each attribute of the class Y_i. and the values of the weights α_s are defined.

Step 4: The values of the attributes $x_{\mathcal{D}}^{s}$are calculated by the formula (3) for the new explored article D, s ∈ K.

Step 5. Determined ν at where $R_\nu = \min\limits_i R_i$, accordingly, the author of the article D is the ν-th author.

5. The results of computer experiments

To test and compare the performance of the abovementioned algorithms, 32 newspaper articles (first sample version) taken from the Internet and 50 (second sample version) newspaper articles in the Azerbaijani language by randomly selected four authors have been included in the database, which are conditionally called *A1, A2, A3, A4.* The authorship identification has been performed respectively by 5 and 8 additional articles written by one of the four authors, the authorship of which is hidden.

32 letters have been used as the attributes in the case of monogram (n = 1), and 835 really possible combinations of letters of the Azerbaijani language - in the case of diagrams (n = 2).

5.1. The experimental results for the study of the first version volume

In the case study, 9 articles of the author A1, 7 articles of the author A2, 6 articles of the author A3 and 10 articles of the author A4 have been reviewed. The total number of letters in each article varied between 3438 and 6859 characters.

For the identification, 1 article $z^1$of the author A1 (3926 letters in the article), 1 article $z^2$of the author A2 (3470 letters in the article), 2 articles $z^3 = (z_4^3, z_2^3)$ of the author A3 (4067 and 4243 letters in the articles) and 1article $z^4$of the author A4 (7463 letters in the article) have been taken.

The first four columns of the Table 1 show the results of the algorithm 1, which uses the attributes based on the monograms. The *j*-th row of the *i*-th column shows the value $R_i(z^j) * 10^3$. Obviously, the correct authorship recognition of the *j*-th article corresponds to the case, where the value of the *j*-th element is the smallest among the elements of the *j*-th row (corresponding values are underlined).

Table 1

	Algorithm 1 (*n*=1)				**Algorithm 3(*n*=1)**				**Algorithm 2(*n*=2)**			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
z^1	***108***	126	124	140	**423**	474	441	494	**164**	341	296	299
z^2	109	**086**	120	075	433	**385**	439	403	408	**238**	295	311
z_1^3	**078**	093	090	118	388	435	**377**	473	339	283	**237**	282
z_2^3	**163**	194	172	233	441	480	**437**	543	412	369	**314**	394
z^4	113	093	134	**083**	432	399	462	**332**	364	304	297	**272**

As seen from Table 1, the authorship of the articles of only the first and fourth authors is correctly identified and the quality of recognition is accounted for 60%.

The second group of four columns of the Table 1 shows the results of the proposed modified algorithm, which is based on the monograms and uses the weights for attributes. As it is seen from Table 1, the efficiency of the recognition significantly improved and equaled to 100%.

The third group of four columns of the Table 1 shows the results of the algorithm 2, which uses the diagrams as the attributes. The obtained recognition results are more stable, i.e. R_ν significantly exceeds for $R_\nu = \min\limits_{i \le i \le n} R_i$, and $j \neq \nu R_j$, while the effectiveness of recognition is 100%.

It should be noted that the features based on the charts for the Russian language first have been used by D.Khmelev. He reviewed the sequence of letters as the realization of a Markov chain. Examining the connection of letters (charts) in the works by 82 authors in Russian, he experimentally showed the most characteristic features of a particular author [19]. The results showed that for the Azerbaijani language, this approach is also more effective. In this case, the authorship of the recognized authors of the papers was determined with the accuracy of 100%.

Table 2 shows the results of checking the stability of letters-monograms, which are based on the analysis of variation of these letters, designed for all the texts of the selected authors. Obviously, the greater the value of variation for any letter is, the less reliable is the use of this feature in the authorship recognition.

In this case, it turned out that the numbers of essential features, which characterize the authors A1, A2, A3 and A4, respectively equal to 18, 20, 20 and 20 and these characteristics vary for separate authors.

The use of modified letter identified the authorship with 100 percent accuracy for all the authors of relevant articles.

In this case, the letters with a high variation (in case study, above 20%) were rejected, which insufficiently characterize the author.

5.2. The results of experiments with the use of the second version volume

13 newspaper articles by the author A1, 11 – by the author A2, 12 – by the author A3 and 14 – by the author A4 have been examined. The total number of letters in the column varies from 3438 to 6859.

Two articles by each presented author have been taken for recognition.

The first four columns of the Table 3 show the results of the performance of the algorithm 1 that uses monogram-based signs. The j-th row of the i-th column shows the value $R_i(z^j) * 10^3$. Apparently, that the correct recognition of the authorship of the j-th article corresponds to the case, where the value of the j-th diagonal element is the lowest among the elements of the j-th row (in the case of accurate recognition of the value $R_i(z^j)$ is denoted by one feature in the bottom, and by one feature on the top in the wrong recognition).

As shown in Table 3, only the authorship of the articles of the fourth author has been correctly identified with the use of the algorithm 1, however, in general the quality of recognition was 50%.

In the second group of four columns of Table 3 shows the results of the proposed modified algorithm based on the monogram, and used for letters of the weights.

As can be seen from Table 3, the detection efficiency significantly improved and reached 62,5%.

The third group of four columns of Table 3 shows the results of the algorithm 2 that uses the charts as the letters.

Table 2

Letter	A1		A2		A3		A4	
	Average evaluation	Variation	Average evaluation	Variation	Average evaluation	Variation	Average evaluation	Variation
A	0.1060	5.54	0.1057	6.46	0.1068	4.09	0.1048	5.34
B	0.0275	***22.06***	0.0293	12.13	0.0327	5.12	0.0216	6.90
C	0.0091	***22.27***	0.0088	16.16	0.0098	***28.12***	0.0104	19.86
D	0.0545	17.49	0.0475	12.45	0.0544	6.50	0.0466	9.77
E	0.0244	16.06	0.0241	8.76	0.0242	16.24	0.0254	16.84
F	0.0058	***26.80***	0.0041	10.62	0.0055	***27.30***	0.0068	***26.75***
G	0.0079	***39.87***	0.0074	***20.48***	0.0080	***27.49***	0.0061	***21.26***
H	0.0144	***30.08***	0.0122	19.38	0.0120	4.12	0.0099	***30.29***
I	0.0380	***22.34***	0.0366	12.34	0.0388	6.56	0.0323	8.51
J	0.0005	***106.4***	0.0005	***39.6***	0.0004	***77.70***	0.0003	***91.57***
K	0.0224	15.84	0.0214	11.98	0.0224	15.04	0.0245	15.57
L	0.0607	14.24	0.0599	8.83	0.0570	8.66	0.0613	5.4822
M	0.0427	15.05	0.0390	15.29	0.0416	7.01	0.0377	7.8337
N	0.0683	15.37	0.0781	7.66	0.0692	8.86	0.0742	10.82
O	0.0240	19.21	0.0203	11.79	0.0216	8.15	0.0174	11.38
P	0.0060	***34.72***	0.0046	***21.03***	0.0063	16.72	0.0062	***29.49***
Q	0.0196	16.88	0.0226	9.20	0.0183	7.16	0.0191	13.73
R	0.0662	15.70	0.0650	7.24	0.0735	1.55	0.0670	4.98
S	0.0309	***21.75***	0.0388	10.92	0.0292	6.86	0.0415	13.87
T	0.0325	19.58	0.0325	10.56	0.0293	3.29	0.0428	9.71
U	0.0271	***21.49***	0.0218	10.26	0.0272	14.87	0.0213	***21.30***
V	0.0123	18.92	0.0129	***22.83***	0.0116	19.42	0.0118	13.16

X	0.0103	***23.42***	0.0072	***30.90***	0.0109	***23.72***	0.0090	***20.05***
Y	0.0320	14.49	0.0328	16.24	0.0324	10.41	0.0335	7.83
Z	0.0178	***24.46***	0.0138	12.70	0.0204	***20.16***	0.0118	13.73
Ç	0.0085	19.08	0.0088	***22.24***	0.0089	12.07	0.0075	17.07
Ö	0.0079	***23.13***	0.0104	***32.95***	0.0090	23.53	0.0080	15.50
Ü	0.0172	18.50	0.0200	***20.90***	0.0215	16.55	0.0166	10.16
Ğ	0.0057	***28.57***	0.0070	17.02	0.0044	***34.08***	0.0057	***28.22***
İ	0.0922	7.03	0.0953	6.18	0.0892	10.33	0.1089	6.32
Ş	0.0147	14.16	0.0129	17.47	0.0156	18.15	0.0119	16.66
Ə	0.0930	6.13	0.0977	4.58	0.0876	3.81	0.0948	7.21

Table 3

	Algorithm1 (*n*=1)				**Algorithm3 (*n*=1)**				**Algorithm2 (*n*=2)**			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
z_1^1	**758**	930	825	1070	**3356**	3631	3625	3895	**2139**	2301	3755	2621
z_2^1	1020	925	**726**	892	3999	3966	**3167**	3607	2150	3684	**1962**	2379
z_1^2	1414	1084	1515	**1215**	4381	**3711**	4700	3770	3844	**2571**	5435	2773
z_2^2	1484	1360	1475	**1308**	4653	**4036**	4225	4155	3813	3666	**2660**	3276
z_1^3	**1495**	1825	1697	2118	**4086**	4634	4965	4987	3471	**2918**	4687	3289
z_2^3	2066	2204	**1909**	1936	4794	5240	**4510**	4981	3606	4214	**2761**	3237
z_1^4	1578	1085	1721	**981**	5005	4035	5123	**3059**	3525	2697	5702	**2223**
z_2^4	1573	1375	1442	**1032**	5641	4516	4616	**3524**	3946	4152	3770	**2674**

As can be seen from Table 3, the use of monogram-based algorithms and the algorithm of the modified monogram method is inefficient for the authorship recognition. The monogram-based algorithm correctly identifies the author only in 50% of cases, and the modified algorithm - in 62,5%of cases.

According to the Table 3, the use of the charts allows us to identify the author with about 62,5% accuracy.

An analysis of the experimental results shown in the Tables 1-3 shows that the recognition results using statistical methods deteriorate with an increase in both the training and examination samples.

The relatively low percentage of authorship recognition in the two versions is mainly explained by the large dimension of the signs space and by the proximity of the signs' values to each other. This is due to the impossibility of separating the set of letters that characterize each author from the letters space with conventional linear hyper-surfaces. On the other hand, it is also associated with a small volume of records in newspaper articles and low informative content of these records.

5.3. The results of applying the support vector machine method

The support vector machine method (SVM - Support vector Machine) was first proposed by Vapnik V.N. [21, 22]. The most popular version of this method has been implemented in LIBSVM package [23]. It should be noted that due to modern developments of many researchers we can

state that the SVM is currently one of the most effective classification (ordering) methods. Unlike the neural networks, the use of a SVM in the multidimensional sings space is more effective, particularly when it is used as *n*-gram letters (diagrams and monograms).

The Gaussian radial basis function (RBF) has been chosen as the core for the support vector:

$$K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}, \gamma \geq 0$$

Numerical experiments have been conducted with the use of LIBSVM package with the letters based on the monogram and diagrams.

Table 4 shows the results of authorship recognition using 1-gram as the signs, and Table 5 shows the results using the signs when *n=2*.

Table 4 (Monogram)

	C=1				**C=10**				**C=100**			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
z_1^1				+	+				+			
z_2^1				+	+				+			
z_1^2				+		+				+		
z_2^2				+		+				+		
z_1^3				+			+				+	
z_2^3				+			+				+	
z_1^4				+				+				+
z_2^4				+				+				+

The version of the second training sample was used for training and recognition. The value of γparameter was taken γ = 1. The value of the parameter C denoted penalty was taken 1, 10 and 100. For small values of the parameter *C* the recognition accuracy was 60-70%.

Note that we have offered above-described modified algorithms that use the monograms and diagrams for the first time. As shown by numerical experiments, the modified *n*-grams have always been more effective than the classic versions of *n*-grams.

Table 5 (Diagram)

	***C*=1**				***C*=10**				***C*=100**			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
z_1^1	+				+				+			
z_2^1	+				+				+			
z_1^2				+		+				+		
z_2^2				+		+				+		
z_1^3	+						+				+	
z_2^3	+						+				+	
z_1^4				+				+				+
z_2^4				+				+				+

Conclusion

The article analyzed the methods of authorship recognition algorithms of the Azerbaijani texts.

The main features were *n*-grams when $n = 1$ and $n = 2$. The recognition algorithms have been built with the use of the statistical approach and support vector method. The recognition objects were the authors of informative newspaper articles, characterized by small size (volume).

The article proposed the modified algorithm of statistical recognition method that uses the extent of *n*-grams variations, which are typical for the articles by the reviewed author.

The article also presented the results of experiments carried out on the recognition of the authors of the articles using the developed algorithms and programs. In addition, it provided the comparison of the obtained results with the results using the support vector method, which has performed more reliable recognition compared to a Bayesian approach.

References

1. Mosteller F., Wallace D.L. Applied Bayesian and Classical Inference, The Case of the Federalist Papers. Reading, MA: Addison-Wesley, 1984, 303p.
2. Burrows J.F. Not unless you ask nicely: the interpretative nexus between analysis and information // Literary Linguist Computing, 1992, vol.7, No.2, pp.91–109.
3. Stamatatos E., Fakotakis N., Kokkinakis G. Automatic Text Categorization in Terms of Genre and Author // Computational Linguistics, 2001, vol. 26, No 4, pp.471–495.
4. Morton A.Q. The Authorship of Greek Prose // Journal of the Royal Statistical Society, Series A, 1965, vol. 128, No 2, pp.169–233.
5. Brainerd B. Weighting Evidence in Language and Literature // A Statistical Approach, University of Toronto Press, 1974, 288p.
6. Holmes D.I. Authorship Attribution // Computers and The Humanities, 1994, vol.28, No 2, pp.87–106.
7. Tweedie F., Baayen H. How Variable may a Constant be Measures of Lexical Richness in Perspective // Computers and The Humanities, 1998, vol.32, no.5, pp.323–352.
8. Stamatatos E., Fakotakis N., Kokkinakis G. Computer-Based Authorship Attribution Without Lexical Measures // Computers and The Humanities, 2001, No 35, pp.193–214.
9. Fürnkranz J. A Study using n-gram Features for Text Categorization, Austrian Research Institute for Artifical Intelligence, 1998, 10 p.
10. Tan C.M., Wang Y.F., Lee C.D. The Use of Bigrams to Enhance // Journal Information Processing and Management, 2002, vol.30, no.4, pp.529–546.
11. Çatal Ç., Erbakırcı K., Erenler Y. Computer-based Authorship Attribution for Turkish Documents / Turkish Symposium on Artificial Intelligence and Neural Networks, 2003, pp. 539–541.
12. Aida-zade K.R., Talibov S.G. Analysis of the effectiveness of the methods of recognition of authorship of texts in the Azerbaijani language // The 5th International Conference on Control and Optimization with Industrial Applications (COIA-2015), 27−29 August, 2015, Baku, Azerbaijan, pp.183.
13. Gasimov S., Ibrahimov I. Analysis of sentences and words used in Azerbaijani texts // The Second International Conference Problems of Cybernetics and Informatics, September 10–12, 2008, Baku, pp. 117–119.
14. Doğan S., Diri B. A New Classification Based on N-grams for Turkish Documents // Author, Type and Gender. Turkish Foundation Union for Computer Science and Engineering Publication, 2010, 3, pp.11–20.
15. Biricik G., Diri B., Sönmez A. A New Method For Attribute Extraction with Application on Text Classification / 5th International Conference on Soft Computing, Computing with Words, ICSCCW, North Cyprus, Famagusta, 2009, pp.4.

16. George H. Estimating Continuous Distributions in Bayesian Classifiers / 11th Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995, pp.338–345.
17. Yasdi M., Diri B. M. Authorship Recognition with Abstract Feature Inference / IEEE 20. Signal Processing and Communication Applications Convention, SIU 2012, Fethiye (18–20 April), 2012, p.4.
18. Orlov Y.N., Osminin K.P. The methods of statistical analysis of literary texts, M .: Editorial URSS / Book House "LIBROKOM", 2012, p.326
19. Khmelev D.V. Text Authorship Recognition using Markov A.A. chains // MGU News, ser.9: Philology, 2000, No2, pp.115–126.
20. Romanov A.S. Text Authorship Recognition Methods based on support vector machine // TUSURReports, No1 (19), part 2, June 2009, pp.36–42.
21. Vapnik V.N. Statistical Learning Theory, New York: Wiley, 1998, 732 p.
22. Vapnik V.N. The nature of statistical learning theory, New York: Springer-Verlag, 2000, 332 p.
23. C.-W. Hsu, C.-C. Chan, C.-J. Lin. A practical guide to support vector classification. // www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
24. Romanov A.S., Mesheryakov R.V. Authorship identification with support vector machine in case of two possible alternatives.www.dialog-21.ru/digests/dialog2009/materials/pdf/67.pdf