The Second International Conference "Problems of Cybernetics and Informatics"
September 10-12, 2008, Baku, Azerbaijan. Section #3 "Modeling and Identification"
www.pci2008.science.az/3/33.pdf

ON THE AUTOMATED TEXT AUTHORS IDENTIFICATION

Vasiliy Makarov[1], Sergey Shako[2], Rinat Kureyev[3]

[1]Technical Institute Branch of Yakut State of University, Russia, *makarov9jku@rambler.ru*
Far Eastern Institute Ministry of Internal Affairs Yakut branch
[2]*shako-77@mail.ru*, [3]*pilot_ka52@rambler.ru*

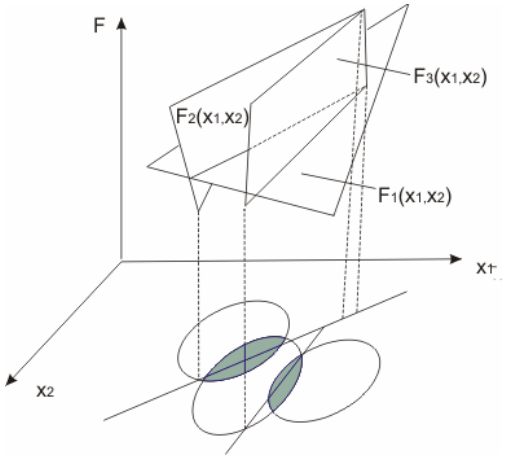
Due to fifteen years' prescription evaluation [1] hundreds of teams work on the problem of the computer-aided identification of texts authors. Ex facte the importance of this research direction ought to decrease because of the avalanche-like spreading of such "masking" facilities as text editors and availability of Internet. However, as it is shown below, the significance of the stated class of tasks can only increase. Thus, there is an appellate judgment of the USA Court of Appeals stating: "the copyright protecting computer programmes should apply to the structure, sequence and organization of the programme, not only to its text" [1, pp. 220 - 221]. That is, the USA Court of Appeals in fact, confirmed the significance of such characteristics as "topographic parameters of the text" in such a specific area of human activity as programming. Similar in sense decisions were made by judicial and legislative bodies in different countries. We would remind you the characteristics of the written text taken by the experts as significant ones [2,3]: 1) Stylistic features; 2) Lexical features; 3) Topographic features.

Stylistic features express the manner of describing – the language and composition of the text. The composition of the text is considered with a view to the presence of such parts as introductory, main and final, the sequence of statement of messages, and the types of the language structures used. In recent years the list was supplemented with "slang" ones which have become widespread in the "advanced" generation medium. Incidentally, the term "advanced" belongs to numerous "slang phrases."

Lexical features characterize the educational and professional level of the authors via their language. The elements of professional turns of speech and international slang are of great interest.

Topographic features characterize the representation of the text (paragraphs, their size, punctuation marks, their location and types (":", hyphen, ";")). Philological and criminalistic analysis experience proved that the stated features are stable and do not much depend on the conditions of execution of the text [2, 4].

In addition to the above mentioned analysis tools, late in 1960s there appeared attempts to use the methods of "linguistic statistics" in criminalistics. The above listed features of the texts **allow to make the task more exact: to formulate the criteria of stability of the written speech in relation to its modifications spread in the Russian Federation by text editors. In terms of informatics and group theory the task explication is possible: to build the system of invariants of the complete text relative to the group of its grammatically acceptable modifications.** The above stated and other terms are explicated in the text. As a working instrument we have chosen easily geometrised method of LDF – linear discriminant functions being of the form: $F = W_1X_1 + W_2X_2 + ... + W_pX_p$. Here {Xp} is the set of identification parameters, {P} is their number, {W} is the weighting factors taking into account the weight of the features. Geometrical sense of LDF becomes transparent if {Xp} and {W} are interpreted as p-measuring vectors in Ep.F(x)=(W, X).



Then after the analysis of the text D_1 decision is made if $F>\alpha$.

For justification of the choice of the diagnostic parameters we used the Zipf- Mandelbrot's law and N.A. Morozov's ideas formulated in a number of works in 1914 – 1915 [4]. N.A. Morozov had predecessors V. Ditterbeger and K. Ritter who studied disputable texts by means of statistic analysis using the most volatile language elements having non-empty classes of synonyms. Essential distinction of N.A. Morozov's method was **the reference thesis: it's not the exclusiveness of the language element that defines the author's style, but the originality in using common language forms which can only be objectively determined statistically**. Further on, *Morozov drew attention to the fact that it's not only necessary to take into account the words having high frequency of application, but to take into consideration that the group of such words is heterogeneous, i.e. they should belong to different parts of speech. He paid special attention to insignificant words, function, or as the author called them, dispositive particles of speech (conjunctions, prepositions, some pronouns and adverbs, etc.).* Answering in the affirmative to the question "Can the authors be recognised by the frequency of such particles as by the features of their portraits?" – Morozov suggests "For that first of all their frequencies should be transferred to the graphs marking each dispositive particle on the horizontal line, and the number of its repetition on the vertical line, and the graphs of different authors should be compared."

The rise in the reliability of the method and authenticity of the results is reached firstly by means of increasing of the accounting language units included in one spectrum, and secondly, by means of increasing of the number of spectra. In the final analysis both the first and the second requirements are fulfilled if the volume of the analyzed text increases. Unfortunately, the lower boundary of the volume of the text has not been established till now.

Zipf-Mandelbrot's law sometimes called "The Law of Organized Message" was gained while solving the task of coding optimization: how can we reduce the total dominant number of letters in the message without losing its sense? It turned out that if the coding is done in the best way (when the most frequently used words are at the same time the shortest ones), words frequencies will be arranged in the following regularity:

$$pi \approx \frac{K}{(B+i)^{\gamma}}; K, B, \gamma - const$$

Thanks to the higher number of constants this formula turned out to be flexible and thoroughly describing the area of the most frequent words in such languages as Russian.

Illustration: Let Z be the volume of the language sample. Then the relative frequency of the most rare word is $p_v = \frac{1}{Z}$, and the total of all the frequencies, defined by Mandelbrot's formula must be unity:

$$\frac{K}{B+1} + \frac{K}{B+2} + ... + \frac{K}{B+i} + ... + \frac{K}{B+v} = 1,$$

on the stipulation that

$$p_1 = \frac{F_1}{Z} = \frac{K}{B+1} \text{ and } p_0 = \frac{1}{Z} = \frac{K}{B+v}$$

It is not too difficult to determine three quantities *K, B* and *v* from these three expressions:

$$K = \frac{1}{\ln F_1}; B == \frac{K}{p_1} - 1$$

However, any lexical sample has always got quite a lot of different rare words each of them found 1, 2 or 3 times. If Mandelbrot's formula is understood literally, the frequencies of all words must be different. To coordinate this frequency area with Mandelbrot's formula, we should admit that the number of different words v_m each of them occurred in the sample with the volume Z units *m* times each is equal:

$$v_m = \frac{v}{m(m+1)}.$$

Relative frequency of the word itself in actual samples can be considered to be constant. That is why the main defining quantity is the volume Z. Should Z increase, the value of K and B decrease, and the vocabulary of the message grows; should Z decrease, the picture is inverse.

The natural question arises: when is Zipf-Mandelbrot's law held? And when might its breach be expected? The formulae given allow the simple evaluation as the problem of constants selection has been solved in them. If we agree to consider that $\gamma = 1$, Z is equal to the actual sample volume and pi is equal to the actually observed frequency of the most frequent word in the sample, thus the theoretical set of frequencies for the given sample lines up explicitly. The result of such comparison found out to be stunning. Mandelbrot's formulae almost always proved to be correct when compared with the frequency data of certain literary works. On the contrary, they almost never described random lexical samples. Large samples pretending to present "the language on the whole" failed completely. In other words: Morozov's spectra method served as an effective tool "in author's hands," but it malfunctioned and gave rise to doubt at the attempt to make its character universal.

At the same time, authorship testing results gained by spectra method do not conform with the results of Mandelbrot's formulae analysis. But another interpretation is possible: these results do not possibly have to conform! They are sure not to correlate with each other, and consequently they cannot be used as diagnostic features in multivariate LDF variant. Namely:

$$F_1 = w_{11}x_1$$
$$F_2 = w_{12}x_1 + w_{22}x_2$$
$$.....$$
$$F_p = w_{1p}x_1 + ... + w_{pp}x_p$$

Or geometrically:

The authors of LDF method at two diagnostic features have created the programme prototype and have tested its efficiency in a range of graduate and even scientific works. The fact of plagiarism was proved with 90% confidence, but not the authentic authorship. The matter is that *the problem of reliability of methods based on using frequency features is not settled as the anonymous texts of such volume are not often found in practice where the main array of such texts consists of much smaller texts in volume.* **The established fact of fundamental unfitness of frequency calculation for attribution of short texts makes us refocus the direction of our search for the adequate showings and prevents us from searching them where they can't occur.** In connection with the stated the interest is attracted to the author's text analysis methods which would:

A) Reflect the author's individual style.
B) Give the authentic conclusions at the message volume of 500 – 1000 wordforms.

As the term "authenticity" presupposes the quantitative evaluation, we had to apply some research in psycholinguistics, namely, the method of "semantic differential" which was developed by American linguist Ch. Osgood. The texts under consideration were analysed by the group of experts according to the specified scales, and the estimates for every scale averaged. Further research [5] showed that in some occasions (fiction, fiction and political text, etc.) eight factors were enough to reach 90% confidence in results, and the procedure of mathematical treatment comes to the rotation of the axes in factors space.

$$a = \frac{1}{2}[(\mu_1, w) + (\mu_2, w)]$$

Summary:

1) Every branch of linguistic activity of social media has invariant texts characteristics.
2) The invariants mentioned can be correctly defined for fixed types of texts only.
3) Two-dimensional LDF variant is enough for the fact of pseudo-authorship establishment.

Literature

1. D. Grouver "Software Protection" Moscow, "Mir", 1992, 286 p.
2. E. Lann "Literature mystification," M-L, 1930, 126 p.
3. L.L. Myasnikov, E.N. Myasnikova "Automatic Recognition of Sound Patterns," "Energiya," 1970, 183 p.
4. Fomyenko "Truth Can Be Calculated", 2008.
5. D. Louly, A. Maxwell "Factor Analysis as a Statistic Method," Moscow, "Mir", 1967, 276 p.