Nadzeya SHUTKO[2], Pavel URBANOVICH[1,2], Pawel ZUKOWSKI[3]

Lublin Catholic University, Poland (1), Belarusian State Technological University (2), Belarus
Lublin University of Technology, Poland (3)

doi:10.15199/48.2018.06.15

A method of syntactic text steganography based on modification of the document-container aprosh

Abstract. *Features of implementation of the text steganography method for the hidden data transmission and protection of intellectual property rights are considered. The method is based on the modification of the spatial-geometric parameter of the container-text - aprosh. Data hiding is performed not only in ordinary, but also in special (soft hyphen, line break, etc.) symbols and spaces.*

Streszczenie. *Przeanalizowano cechy wdrożenia metody steganografii tekstowej w celu ukrytej transmisji danych i ochrony praw własności intelektualnej. Metoda oparta jest na modyfikacji parametru przestrzenno-geometrycznego tekstu-kontenera - aprosh. Ukrywanie danych odbywa się nie tylko w zwyczajnych, ale również w specjalnych (mijany łącznik, przerwa liniowa i td.) symbolach i spacjach.* ***(Metoda syntaktycznej steganografii tekstowej bazowana na modyfikacji aproszu dokumentu-kontenera****).*

Keywords: *text steganography, property rights, aprosh, steganography strength*.
Słowa kluczowe: *steganografia tekstowa, prawo własności, aprosz, odporność steganograficzna*.

Introduction

The development of information technology has led to the fact that now they put the information on a par with the generally accepted material values. Access to it with the advent of global computer networks has become incredibly simple. Simplicity and speed of such access significantly increased both the threat of data security breach, as well as the threat of unauthorized access (without author's permission) to information.

The main objects of copyright relating to information technologies, including publishing technologies, are: paper and electronic versions of various text or other documents, databases, computer programs. This means that Internet objects also refer to intellectual property. The logic of regulating the Internet itself and the relations connected with the use of its resources and capabilities should be in the plane of both national and international law.

Thus, the problem of protecting information and protecting copyright for text documents is becoming increasingly important.

A method of solving of this problem on the basis of text steganography is investigated and analyzed in some articles, for example [1]. The specifics of algorithmic and software implementation of the steganographic methods that are based to protect of digital documents against unauthorized use are analyzed in [2].

Contrary to cryptography which purpose is hiding data by encrypting them, the purpose of steganography is to hide the fact of the transfer of confidential messages. It is because of steganography system of protection is achieved the greatest degree of resistance to intentional attacks to destroy or to identify hidden information. The steganographic system (stegosystem or steganosystem) – a set of tools and techniques that are used to form a secret channel of information transfer [3-5].

The steganosystem forms the channel, that carries the filled container. This channel is considered to be exposed to the influence from the violators.

Embedding messages can be performed using the key, and without its use in the steganographic system. To increase the steganoresistance of system the key can be used as a verification tool. It can also have an impact on the distribution of bits of the message within the container on the order of forming a sequence of embedded bits of messages.

In text steganography the secret information is embedded in the document-container on the basis of the modification of the parameters of the text characters.

In this paper we analyze some important aspects of the practical implementation of the method based on the modification of such a spatial-geometric font parameter as aprosh (letterspace).

Aprosh and others spatial-geometric font parameters

Fonts of texts, generated and processed by word processors (for example, MS Word) are characterized by approximately 20 basic space-geometric parameters. In addition to aprosh among them also are: em-square, point size, leadind and others.

Information that modifies the original font settings for the text-container can be used, if necessary, to prove the intellectual property right. Thus, this modification of the text parameters is associated with some formal transformation or encoding of the text symbols. Or more precisely: in our case, the encoding of a symbol is a change in the detailed feature of this symbol.

As already mentioned above, we modify the aprosh.

The following circumstance should be taken into account: a fundamental feature of fonts is the separation of information about the form of characters from the process of their reproduction on the raster output device. If the profiles of font characters can be described by a variety of ways, then the task of the playback, in the final analysis, boils down to activating some points: flashing the pixels on the display screen or filling them with ink when printing on the printer. It is easy to see that in the reverse transformation (extraction of a message) ambiguities may arise due to the font structure (headset).

Such uncertainty can be caused, for example, by a violation of the symmetry of some characters of the text (for example, by the occurrence of different distances between vertical strokes of the letter "Ш" in the belarusian or russian alphabet), which sharply distorts their shape and makes it difficult to restore the author's information.

Another feature is the reproduction of characters on devices with low resolution (up to thousands of pixels per inch), especially when text is printed with a small size (12 or less), which leads to scaling errors. Scaling occurs in absolute coordinates relatively to some arbitrary point and always leads to an integer result. This raises the problem of rounding off non-integer results. For example, if the coordinates of some element of the symbol in the

coordinate system of the contour description are equal to (200; 100), then when the contour size is reduced 3 times, they are transformed into (66,66; 33,33). Since we need integer values, they will turn into (67; 33), that is, the value of the horizontal coordinate will increase slightly (by one third of the pixel), and the vertical coordinate will decrease by the same amount.

Aprosh – the spacing between neighboring letters or other font symbols. According to the existing technical rules the normal aprosh should be equal to half the font size. The parameter being analyzed determines such an important characteristic of the text as its readability.

The above features of the aprosh are a good prerequisite and the basis for using the aprosh as a parameter that can be modified when information is embedding in a text-container.

The concept of used model of steganographic system

In general, the steganographic system Σ formally is described by the formula of form:

$$\Sigma = (F, F^{-1}, M, C, K, S), \tag{1}$$

where M, C, K, S – respectively the elements of the sets of messages, containers, keys, steganoconatainers (steganomessages – container with embedded message), as well as the transformations that connect them: F – transformation of the container by the embedding of message into it based on key information, which result is a steganoconatainers; F^{-1} – corresponds to the process of extracting the embedded message from the steganocontainer.

We will define the abstract steganographic system as a set of mappings of a single space (the set of possible messages, M) to another dimension (the set of possible stegomessages, S). Or vice versa.

Here are the characteristics of the models [6]:

1) the processes of embedding/extraction of information, which are based on the corresponding basic algorithms, from a formal point of view are defined by the types of the embedded/extracted information, container and a selection of specific container elements or groups of these elements used to accommodate the relevant message components; such basic algorithms used for the textual stenography can for example be: methods of Line-Shift Coding, Word-Shift Coding and others [7-8];

2) the basis of the method and the corresponding methods we put essential space-geometric and color characteristics of the basic elements of a text container (or else – fonts);

3) an important distinctive feature of the mathematical model is the identification if the selected stenographic method (based on the modification of the specific space-geometric of color setting characters of text) with essential process information; in our opinion, for the unauthorized user the information remains secret;

4) other hidden parameters of the stenographic process we will consider as additional key information.

We will base the model using the following notation and regulations. Let M be a finite set of messages that can be hidden in the container: $M = \{m_1, m_2, \ldots, m_n\}$; in the context of solved in the framework of the thesis, the tasks specified messages are text documents.

C – is the finite set of all admissible container (cache files or text cache documents): $C = \{c_1, c_2, \ldots, c_p\}$, general case $p > n$;

K – set of keys, generally we will understand methods and deposition message algorithms in container or other operations preliminary transformed embedded message or selecting the elements in container for such a deposition: $K = \{k_1, k_2, \ldots, k_z\}$.

An arbitrary hidden message m_i ($1 \le i \le n$) can be hidden in the container c_j ($1 \le j \le p$) using k_m ($1 \le m \le z$) key: $m_i \in M$, $c_j \in C$, $k_m \in K$. The result of this type of transformation is a full container (or steganomessage) s_q, pertaining to a set of full container or steganomessages S: $S = \{s_1, s_2, \ldots, s_r\}$.

The function F, defined by $M \times C \times K$ with the values in S, will be identified with deposition or insertion of messages m_i from the set M in container c_j in the set C on the basis of key of set K, stipulation the use of an appropriate algorithm deposition and space (aprosh or other) parameters of container c_j set C:

$$F: M \times C \times K \to S. \tag{2}$$

The function F^{-1}, specific for $S \times K^*$ ($K^* = \{k^*_1, k^*_2, \ldots, k^*_z\}$, for general case $k_m \neq k^*_m$; $k_m \in K$, $k^*_m \in K^*$) with the values in M, we will identify with the recovery of the hidden message $m_i \in M$ from steganomessage $s_q \in S$ ($1 \le g \le r$):

$$F^{-1}: S \times K^* \to M, C. \tag{3}$$

The expression (3) formally describes the procedure of extraction a message from a container using the same chosen method. Accordingly, each concrete mapping $(F^{-1})_w$, where $w = 1, 2, \ldots, l$, of the plurality of F^{-1} corresponds to a particular algorithm or method of embedding information m_i to container c_j using a specific key k^*_w.

The set K mentioned in (1) can be considered consisting of a certain number of subsets. One of these subsets contains the keys K^0, which determine all possible methods of embedding/ extraction of information. In particular, one of the keys of this subset determines the steganographic method we are considering. Another subset of the keys, K^1, can be correlated with the transformation of the embedded message to enhance the steganographic (by analogy with the cryptographic) persistence of the transformation. Such transformations can be, for example, cryptographic transformation or transformation based on redundant coding, or a combination thereof, or other combinations of transformations. And, finally, it is possible to allocate one more subset of the keys, K^2 – it determines the order of the selection of the container symbols for the embedding of the next bit (or several bits) of the secret message.

Thus, our set K can be defined as consisting of three subsets:

$$K = \{K^0, K^1, K^2\}. \tag{5}$$

More details the features of modeling and analysis of multi-key steganographic systems are described in [9].

A formal description of a method based on the aprosh modification

The idea of the proposed method is as follows.

Embedding a message into a container can be based on modifying the base (defined by the word processor by default) value of a aprosh, a_o, by changing it from the base to some maximum, a_{max} (or minim, a_{min}), which should not be visually different from the standard one. Such a change is made with a certain step (discretely) Δa_t, each value of which is assigned a certain bit or a certain combination of bits that are a part of the embedded message.

Change of aprosh value between two definite symbols of the text (container) relative to the base value of a_0 by a small distance (points (pt) or parts of a point) can be formally represented in the following form:

$$a'_t = (a_0 + \Delta a_t). \tag{6}$$

Such a change should not cause a visually noticeable compaction (Δa_t <0) or a vacuum (Δa_t> 0) of symbol groups. In the word processor MS Word the aprosh can take values in the range from 0 to 1584 points.

For an example and a visual representation of the features of setting a given spatial-geometric font parameter in figure 1 a text line with various parameters of an aprosh is shown.

При использовании метода осаждение секретного
При использовании метода осаждение секретного
При использовании метода осаждение секретного
При использовании метода осаждение секретного
При использовании метода осаждение секретного
При использовании метода осаждение секретного сообщения

Fig.1. Examples of using a different size of aprosh

In this example the second line is framed using a standard (regular) aprosh. In the first line compaction is applied in all words except the first: in the second word – by 0.1 pt, in the third – by 0.2 pt, in the fourth – by 0.3 pt, in the fifth – by 0.4 pt, in the fifth – by 0.5 pt; in the third line words from the second to the fifth are formed with a change in Δa (see (6)) in the opposite direction, i.e with a rarefaction.

In the fourth line, condensation (negative Δa) or rarefaction (positive Δa) was applied only to individual symbols of the first ("При") and second ("использовании") words: for "П" – Δa = -0.1 pt; "ри" – Δa = -0.2; "и" – Δa = 0; "с" – Δa = 0.1; "н" – Δa = 0; "о" – Δa = -0.1; "л" – Δa = -0.2; "ь" – Δa = -0.3; "з" – Δa = 0.4; "ов" – Δa = -0.3; "ан" – Δa = 0.2; "ии" – Δa = 0 pt. The fifth line is completely formatted at Δa = 1 pt, and the sixth one at Δa = -1 pt.

As can be seen from this example, using a different aprosh value (better and positive and negative) from 0,1 to about 0,5 pt without a careful analysis is impossible to notice visually.

A feature of this method is the possibility of a one-time placement (in an aprosh of one character) of the number of bits determined by the discrete difference between the minimum and maximum values of Δa. For example, if you count from Δa_{min} to the set interval Δa_t in the form of a parameter $0.1 \cdot n_e$ (pt), then the number of conditional discrete units n_e represented in binary form, determines the number of bits that can be placed in container; for example we use Δa_{min} = -0.5 pt, and Δa_t = 0.3 pt.

The difference between these values is 0.8 pt: 8·0.1 or $(n_e)_t$ = 8 (in binary it is 1000; in the first approximation, such a binary combination can be placed (embedded) by modifying a specific aprosh). On this basis, various variants of the encoding of the embedded combinations (message M) can be developed.

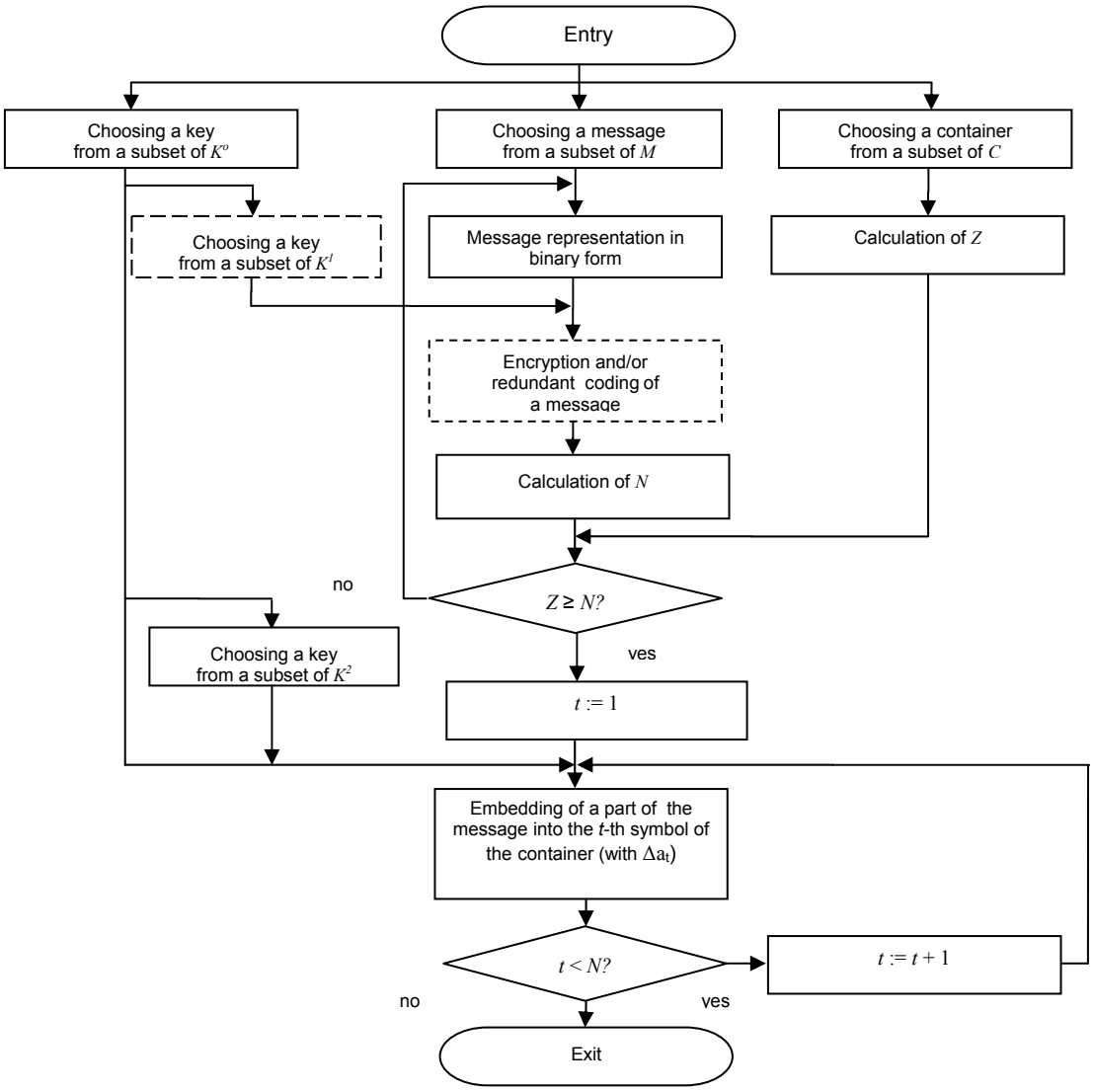

Fig.2. Algorithm flow diagram of a secret message embedding into a text-container based on modification of the document-container aprosh

Features of the algorithm implementation

The general structure of the algorithm implementation of steganography methods based on the modification of the spatial-geometric and color parameters of the text-container is given in [2].

With more detailed analysis of the algorithm of the method we are considering it should be taken into account the following features. The maximum possible volume N (in bits) of the embedded message depends mainly on two parameters: the volume of the text-container, Z, and Δa size. The aprosh size, $a_t^{'}$, of the modified symbol (i.e., the symbol with the embedded information) is calculated on the basis of the formula (6) – it is formed from the aprosh value Δa_t of the same symbol at the document-container and specified in the text processor settings (via special software) of the bias Δa_t, where t = 1 ... Z.

The embedding procedure is performed cyclically N times.

Note also that the selection of text symbols for deposition can be performed based on one of the key information parameters, K^2:

1) local (in order) choice;
2) random (pseudorandom) choice;
3) random (pseudo-random) choice with memory (choice of the next element depends on the position of the previous element).

Figure 2 provides an algorithm flow diagram of a secret message embedding into a text-container by the method.

Operations on the embedded information, determined by the blocks of the algorithm, which are marked with dashed lines in figure 2, are optional.

The inverse transformation algorithm (information extraction) involves performing basic operations (after initializing the container and key information) in a sequence back to the sequence in relation to the information embedding algorithm.

As follows from the analysis of these algorithms, they belong to the class of complexity $O(n)$.

Comparative evaluation of the effectiveness of the method and conclusions

In [2] it is shown that the steganographic method based on the modification of the apropos is one of the most effective in terms of the maximum volume of the embedded information. Another important parameter characterizing the effectiveness of the method is the steganocontainer's strength to attacks based on visual analysis. We conducted an experiment. Students and schoolchildren were handed out text documents (steganocontainers), in which some characters were modified by embedding in them from 1 to 6 bits of a secret message. The results of the experiment are shown in table 1.

Table 1. The results of the experiment

Number of built-in bits per text symbol	Negative reply, %	
	Students	Schoolchildren
1	100,00	100,00
2	90,40	97,50
3	9,60	2,50
4	0,00	0,00
5	0,00	0,00

A negative reply means that the respondent did not notice any differences in the structure (format) of the text document in comparison with the standard ones. The received digital indicators can be considered as one of possible subjective estimations of enough good strength of a method to attacks on the basis of the visual analysis.

Authors: *dr Nadzeya Shutko, Belarusian State Technological University, 13a, Sverdlova Str., 200050 Minsk, Belarus, E-mail: shutko_bstu@mail.ru; prof., dr hab. Pavel Urbanovich, Belarusian State Technological University; Lublin Catholic University, 14, Racławickie Ave., 20-950 Lublin, Poland, E-mail: pav.urb@yandex.by; prof. dr hab. Pawel Zukowski, prof. PL, Lublin University of Technology, 38a, Nadbystrzycka Str., 20-618 Lublin, E-mail: p.zhukowski@pollub.pl*

REFERENCES

[1] Urbanovich P., Chourikov K., Rimorev A., Urbanovich N., Urbanovich N., Text steganography application for protection and transfer of the information, *Przegląd Elektrotechniczny*, 86 (2010), nr.10, 95-97.

[2] Shutko N., The use of aprosh and kerning in text steganography, *Przegląd Elektrotechniczny*, 92 (2016), nr.10, 222-225.

[3] Konahovich G.F., Puzyrenko A.U., *Computer steganography*, Kiev, MK-Press, 2006 (in Russian).

[4] Gribunin V. G., Okov I.N., Turincev I.V., Digital steganography, Moscow, Solon-Press, 2002, 272 (in Russian).

[5] Urbanovich N., Development, analysis of efficiency and performance in an electronic textbook methods of text steganography, Printing future days: 4th International Scientific Conference on Printing and Media Technology, Chemnitz, 2011, p. 189-193.

[6] Shutko N.P., Romanenko D.M., Urbanovich P.P., Mathematical model of the text steganography on the base of modifying the spatial and color settings of text characters, Proceedings of BSTU: Physics and mathematics. Informatics, Minsk, 179 (2015), nr. 6, 152–157.

[7] Brassil J., Low S., Maxemchuk N.F., O'Gorman L., Electronic Marking and Identification Techniques to Discourage Document Copying, *IEEE Journal on Sel. Areas in Commun*., 13 (1995), nr. 8, 1495-1504.

[8] Low S.H., Maxemchuk N.F., Lapone A.M., Document Identification for Copyright Protection Using Centroid Detection, *IEEE Trans on Commun*., 46 (1998), nr.3, 372-381.

[9] Urbanovich P., Shutko N., Theoretical Model of a Multi-Key Steganography System, in: Recent Developments in Mathematics and Informatics, Contemporary Mathematics and Computer Science Vol. 2, Ed. A. Zapała, Wydawnictwo KUL, Lublin, 2016, Part II, Chapter 11, pp. 181-202.