Nadya SHUTKO

Belarussian State Technological University

doi:10.15199/48.2016.10.51

The use of aprosh and kerning in text steganography

Abstract. *The specifics of algorithmic and software implementation of the steganographic methods that are used to protect of digital documents against unauthorized use are analyzed in this paper. The secret information is embedded in the document-container on the basis of the modification of the parameters of the text characters, such as aprosh and kerning. The efficiency of these methods in comparison with some known methods of syntactic text steganography is analyzed.*

Streszczenie. *W artykule przeanalizowano specyfiki algorytmiczne i programowe wdrażania metod steganograficznych, które są wykorzystywane do zabezpieczenia dokumentów cyfrowych przed użyciem nieuprawnionym. Tajna informacja ukrywa się w dokumentcie-kontenerze na podstawie modyfikacji parametrów znaków tekstowych, takich jak aprosz i kerning. Została przeanalizowana skuteczność tych metod w porównaniu z niektórymi znanymi metodami syntaktycznej steganografii tekstowej.* ***(Wykorzystanie parametrów aprosza i par kerninga w stenografii tekstowej)****.*

Keywords: copyright, text steganography, aprosh, kerning pairs.
Słowa kluczowe: prawo autorskie, stenografia tekstowa, aprosz, pary kerninga.

Introduction

In recent years' text documents, codes of the computer programs have increasingly become the objects that are used by others for commercial purposes without permission of the authors. Thus, the problem of definition and proof of ownership of the various documents relating to the field of information technology is becoming increasingly important.

One of the ways to solve this problem is the use of steganography. In our case — the text steganography, because the secret information (stegomessage) hides in the text document (container). The resulting document-container with hidden information will be called stego.

Two classes of text steganography methods are known: syntactic and linguistic. The syntactic methods do not influence the semantics of the text. The linguistic methods are based on equivalent transformation of the text files, that preserve the semantic content of the text. To the suggested and researched syntactic methods we include the following three known methods [1]:

- line-shift coding (the change in the difference between the lines of the text),
- word-shift coding (the change on the distance between the words in the text),
- feature coding (making the specific changes of fonts in some individual letters).

The paper examines the characteristics of some algorithms that implement the methods of text steganography to solve this problem. These methods are based on the using the geometric parameters of the font: aprosh and kerning.

The main characteristic of the methods is that the private data (that is hiding inside the text) does not alter the logical nor the semantic nature of the text. However, this data will remain invisible for the other users. At the same time, they will allow (with the use of extraction and decryption) to get the copyright of the document.

The proposed methods

Aprosh — the spacing between neighboring letters or other font symbols. In the text documents there are such combinations of characters that form a visual hole or condensation (for example, in the texts based on Cyrillic — these combinations: ГА, ТА, АТА, ЬТ, AW and so. etc., based on the Latin alphabet — AY, Av, T, ff, and on the basis of the Greek alphabet — ΘA, ΔO, λκ). Visual alignment of interalphabetic spaces in such combinations is called kerning. Below there are examples of the use of kerning pairs for different characters (Fig. 1).

As can be seen from the figure below, the distance between the characters is not the same. Obviously, the spacing between A and W is much greater than between W and E (in the first case). The visual perception of the text has improved after applying the automatic (in the second case) and manual (in the third case) kerning.

The number of kerning pairs for different fonts is not the same. For example, there are 909 kerning pairs for the font Arial (which is written this work).Information about kerning tables for different pairs of fonts were received with usage [2].

Fig.1. The use of kerning

Table 1 shows the comparison of the characteristics of the number of kerning pairs in some fonts. Four basic fonts (the most frequently used in the word processor Microsoft Word) were taken for the analysis: Arial, Times New Roman, Calibri Regular, Cambria.

Table 1. The kerning pairs

The font name	The number of kerning pairs
Times New Roman	867
Arial	909
Calibri	26 706
Cambria	29 715

The method of text steganography, based on the change of kerning, allows to precipitate secret information in the document container. This information can be used to prove the ownership of the document.

When using this method, the embedding of the stegomessage into the text document-container is made by modifying aprosh from the basic and up to the maximal/minimal value. These values do not differ visually from the standard, with the use of s specific increment,

each of which is assigned a value of a specific combination of bits.

This selective choice allows to compensate for the unevenness of the visual density of the text, obtained using the basic aprosh for each letter. In order to embed the stegomessage into the container there need to be used (character of the stegomessage encoding «1» or not to use kerning (character of the stegomessage encoding «0») for certain characters.

Further analyze the content ratio of the number of characters contained in the stegomessage (N_m) to the total number of the characters in the container (a text of document, N_c).

$$P = \frac{N_m}{N_c} \times 100\% \,. \tag{1}$$

Next, a comparative analysis of the number of kerning pairs, which are found in the same text document created on the basis of various groups of fonts, will be made.

The text from [3] will be used as the container. The total number of characters with spaces in the document is 2 758. The secret message will be the author's name in Russian and the name of the university — «Шутько Надежда, Белорусский государственный технологический университет». The number of characters with spaces of the stegomessage — 71 (it's approximately 3% of the total number of characters in the document-container).

To determine the efficacy and feasibility of the proposed method (changing kerning) it is necessary to analyze the document-container in the presence of a kerning pairs. We developed a special software tool («Kern») for this analysis. Its work is based on the use of underlying tables of kerning pairs in a variety of fonts. Analyzing the desired text, the software searches for coincidence in combinations of letters with the base tables. Then it displays the kern pairs, which are met in the text (how many times they were found), and the total number of kerning pairs in the document. Figure 2 shows the main window of the software.

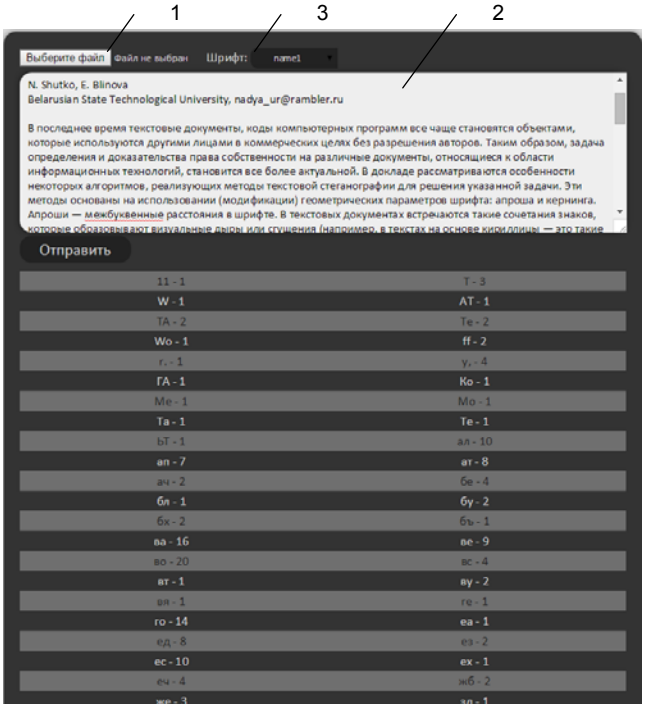


Fig.2. The main window of the software tools «Kern»

The required text document can be selected by pressing the button «Select File» (1), located in the upper left corner of software window, or you can enter the text directly into a text input field (2). The analyzed text document can be written in both Russian and any other language. In the area «Font» (3) you can select the font of your document to correctly parse the text in the presence of a kerning pairs (due to every group of fonts having its own table of kerning pairs).

The total number of kerning pairs was determined in a document-container with the help of software tools «Kern»: in this case — 461 (it is about 34 % of the total number of characters in the document-container). These comparative results are shown in the table 2.

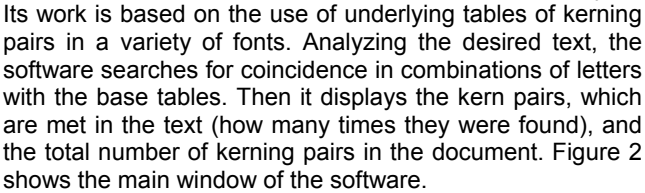
Table 2. The number of kerning pairs in the document-container

The font name	The number of kerning pairs
Times New Roman	457
Arial	461
Calibri	399
Cambria	300

From the analysis of the obtained data it can be concluded that for more effective and less visually noticeable embedding secret information is better suited the text document written by font Arial, the font Cambria is the least appropriate.

The software tool known from [4] was used to embed a secret message in the container document. This software also allows to embed stegomessage (any file) into the container (a test document in the format .docx) by selected by the user algorithm.

By varying the number of bits allocated for the distance between the characters (aprosh) and kerning, we observed the changing of the symbols in the document-container. A good examples of this method are presented in Figures 3-6.


Использование апроша и кернинга в текстовой стеганографии

N. Shutko, E. Blinova
Belarusian State Technological University, nadya_ur@rambler.ru

В последнее время текстовые документы, коды компьютерных программ все чаще становятся объектами, которые используются другими лицами в коммерческих целях без разрешения авторов.Таким образом, задача определения

Fig.3. Example of the deposition information in kerning when 1 bit is modified (here shows a part of the filled container)

Использование апроша и кернинга в текстовой стеганографии

N. Shutko, E. Blinova
Belarusian State Technological University, nadya_ur@rambler.ru

В последнее время текстовые документы, коды компьютерных программ все чаще становятся объектами, которые используются другими лицами в коммерческих целях без разрешения авторов.Таким образом, задача определения

Fig.4. Example of the deposition information in kerning when 8 bit is modified

Использование апроша и кернинга в текстовой стеганографии

N. Shutko, E. Blinova
Belarusian State Technological University, nadya_ur@rambler.ru

В последнее время текстовые документы, коды компьютерных программ все чаще становятся объектами, которые используются другими лицами в коммерческих целях без разрешения авторов.Таким образом, задача определения и

Fig.5. Example of the deposition of information in aprosh when 1 bit is modified (here shown a part of the filled container)

Использование апроша и кернинга в текстовой стеганографии

N. Shutko, E. Blinova
Belarusian State Technological University, nadya_ur@rambler.ru

В последнее время текстовые документы, коды компьютерных программ все чаще становятся объектами, которые используются другими лицами в коммерческих целях без разрешения авторов.Таким образом, задача определения

Fig.6. Example of the deposition of information in aprosh when 8 bit is modified (here shows a part of the filled container)

The embedding function realizes and adds the text steganography algorithms taking into account the opportunities of text processor Microsoft Office Word 2007. The data hiding is done not only in the kerning pairs of characters but also in the special (hyphens, line break, etc.) characters and spaces.

The text processor allows to indicate the kerning of the certain size for each character of the document (this option stores the number from 0 to 1638 points). The hiding (precipitation the author information) is made by changing the kerning of characters in the range from 0 to 1023 points, which provides the placement of up to 10 bits of data in each sign of the text.

The table 3 contains information about the effectiveness of embedding stegomessage in the document-container. We have used the different algorithms for embedding: the number of embedded bits per symbol of the text document (1 to 10) was changing. The experiment showed that even with a closer examination of the text the modification 1 to 4 low-order bits that define the character spacing (aprosh) is unnoticed [4].

As can be seen from the table 3, the number of characters with the data (with a modified kerning; column 3) decreases in proportion to the number of data bits precipitated by increasing the amount the embedded bits per character of the text (column 1). For example, in the first line of the table the initial value of number of characters equals the number of data bits of the embedded message (568). Understandable, the greater the number of bits you need to embed a symbol, the clearer it will be embedded in the text of the private message.

Table 3. Embedding the stegomessage

The number of bit per symbol	Stego, bit	The number of letters with the data
1	2 769	568
2	5 538	284
3	8 307	190
4	11 076	142
5	*13 845*	*114*
6	*16 614*	*95*
7	*19 382*	*82*
8	*22 152*	*71*
9	*24 921*	*64*
10	*27 690*	*57*

During the embedding, the stegomessage is represented in a binary code in the form of a special array containing the complete information necessary to restore the data from the container document

The structure of this array is represented in figure 3.

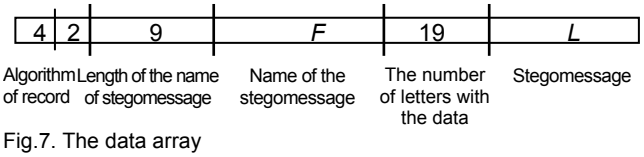


Fig.7. The data array

At first in this array the algorithm of embedding the stegomessage into the container is encoded. This parameter takes 6 bits — 4 bits are for storage of the kerning values (from 0 to 10 in the binary representation), two bits — for aprosh (from 0 to 3 in the binary representation).

The stegomessage can be represented as files of several formats (.doc, .txt, .bmp), that is why when you remove it from the container not only it's name must be known. Therefore, the name of stegomessage is located further in the array of bits. For this purpose, 9 bits, which keep the length of the filename stegomessage (the length of file name must be less than 512 characters), are reserved.

The file name of the stegomessage occupies *F* bits in the container, which are calculated according to the formula:

$$F=N_S \cdot 16, \tag{2}$$

Ns is the length of file name; 16 — constant, the file name is presented in a UNICODE format, that takes 2 bytes (16 bit) per 1 symbol.

Representation of a file name in this format allows to store this information with the characters of almost all written languages, as well as the use of special symbols (©, ®, etc.). Further, the array of bits contains the data about the number of characters of text containing embedded data. This entry contains 19 bits, because the maximum size of stegomessage file is 512 KB. The worst algorithm of embedding is the cover-up in 1characters of text 1 bit of data.

Finally the file of stegomessage size of *L* bit is located in the array.

The algorithm of embedding the stegomessage in a text document-container, the size of which is *N* characters of the alphabet, is shown below.

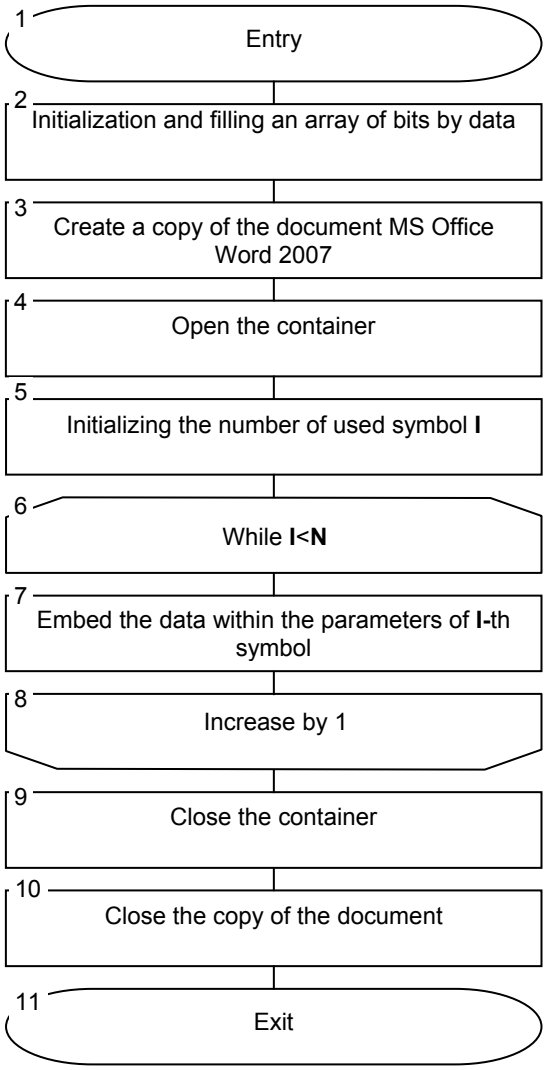


Fig.8. The algorithm of the use kerning during embedding information

After filling in and initialization of the array of bits the text editor Microsoft Office Word 2007 launches (an instance of the document is creating), then the container file opens in it. Embedding data takes place by sequential changing the value of kerning for each character of the document, according to the array of bits. The operations corresponding to blocks of the algorithm from 4 to 7 are executed *L* times.

After completing the process of embedding the data, the changes of the container are being saved and then the text processor Microsoft Office Word 2007 is closed (closing of instance of the document).

The general scheme of the data conversion is shown in figure 9.

An array of bits after the process of the embedding is contained in the parameters of symbol of the file container.

Discussed features of realization of the method, as can be seen relate to letter spacing distance (aprosh) any pair of characters of the document container. If these pairs are the mentioned above kern pairs, then the algorithm is complemented by appropriate search operations and memorization of their location in the container. Formally, it gives additional possibilities in comparison with the change of aprosh.

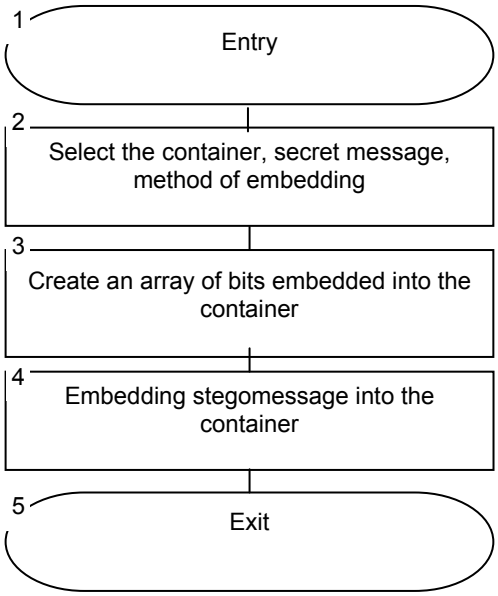


Fig.9. The scheme of data transformation

Conclusion

In this paper we looked at the peculiarities of the algorithmic and program implementation of the stenographic methods, which are used for the security of the digital documents against unauthorized use. The proposed by the author, steganographic methods of protection of copyright on text documents are based on the changing of the spatial parameters text symbols of the container. The hidden information is deposited into the document-container with the use of modifications of such parameters as aprosh and kerning. The effectiveness of such methods is considered and given a comparative evaluation of the existing methods of syntactic text steganography.

There is the algorithm of embedding secret information (stegomessage) into a text document (container) in the paper.

The main disadvantage of the syntactic methods is that they are not resistant against attacks such as print-it-out-type-it-in, and file container can be broken by finding patterns used in the text.

The methods of using the aprosh and kerning of the text document container were considered. These parameters are the modifiable parameters in during the deposition the secret (the author) information, they can be used either singly or in combination. In the second case, it is increasing the maximum volume of stegomessage which can be placed in the container. It is obvious that these methods are more effective in comparison to other known methods of syntactic text steganography. For a quantitative estimation of the analyzed comparison we have taken an electronic version the book [5] as a document container, containing 7 843 742 characters. The final results are presented in Table 3.

Table 3. The number of kerning pairs in the document-container

Methods	Stegosigns	The density of filling, %
Line-shift coding	40 553	0,517
Word-shift coding	1 111 979	14,177
The methods of changing aprosh, kerning	7 843 742	100,000

Authors: *Post Graduate Student Nadzeya Shutko, Belarusian State Technological University, 13a, Sverdlova Str. 220000 Minsk; E-mail: nadya_ur@rambler.ru.*

REFERENCES

[1] Maxemchuk N., Low S., Brassil J. T., O'Gorman L., Document Marking and Identification using Both Line and Word Shifting. Infocom '95, Boston, Mass. Apr 4-6, 1995, 853-860.

[2] FontExpert 2014 12.0 [Electronic resource]. Access mode: http://fontexpert-2014.software.informer.com/12.0/. Title screen, (Reference data: 14.05.2015)

[3] Shutko N., Blinova E., The use of aprosh and kerning in text steganography/ Proc. of Int. Conf. on New Electrical and Electronic Technologies and their Industrial Implementation, Zakopane. (2015), 77

[4] Urbanovich P., Urbanovich N., Chourikov K., Rimorev A., Niektóre aspekty zastosowania metod steganograficznych do przechowywania powiadomień tekstowych, Przeglad elektrotechniczny. Warszawa, (2010) n.7, 95-97

[5] Ozhegov, S. I. The Explanatory Dictionary of the Russian Language, 2nd ed. M.: AST, (2010), 736 p.

Authors: *Post Graduate Student Nadzeya Shutko, Belarusian State Technological University, 13a, Sverdlova Str. 220000 Minsk; E-mail: nadya_ur@rambler.ru.*