Pavel URBANOWICH[1,2], Konstantin CHOURIKOV[2], Andrey RIMOREV[2], Nadzeya URBANOVICH[2]

Catholic University of Lublin (1), Belarusian State University of Technology (2)

Text steganography application for protection and transfer of the information

Abstract. *In the paper some aspects of steganography methods application of the latent accommodation of one text messages in others text messages named by containers are considered. The comparative analysis and an estimation of linguistic and syntactic methods of text steganography are performed.*

Streszczenie. *W artykule zostały przeanalizowane niektóre aspekty zastosowania metod steganograficznych do przechowywania powiadomień tekstowych w innych powadomieniach tekstowych. Przeprowadzono również analizę porównawczą metod lingwistycznych i syntaktycznych w steganografii tekstowej.*

Keywords: steganography, text message, container.
Słowa kluczowe: steganografia, powiadomienie tekstowe, kontener.

Introduction

In the last few years there was a fast growing interest in ways in methods of hidding some information in other information [1, 2]. The fact that an unlimited number of perfect copies can be illegally produced led people to study ways of embedding hidden copyright information and serial numbers in audio and video data. The concern that privacy could be eroded led to anonymous remailers, techniques for making mobile computers harder for the third party to trace. The restrictions of some governments concerning the availability of encryption services motivated people to study and find methods for communicating secretly. Those are new techniques that the scientists are currently developing and who are in permanent improvement. The science which studies methods of concealment of one information in other information is called steganography.

In traditional steganography the set-up is formulated as a prisoner’s problem: Alice wishes to send a secret message to Bob by hiding information in a cover message (container). The stego message (cover+secret message) passes through Wendy (a warden) who inspects it to determine if there is anything suspicious about it.

Applications of steganography include covert communications, watermarking and fingerprinting that seem to hold promise for copyright protection, tracing source of illegal copies, etc. The use of a stego key may be employed for encryption of the hidden message and/or for randomization within the stego scheme. The process may be defined as follows: *cover medium* + *hidden information* + *stegokey = stego-medium.*

There are several issues to be considered when studying steganographic systems. One among the key performance measures used to compare different message embedding algorithms is steganography capacity [3].

The interrelation of known steganography methods is presented on Fig.1 [4].

Technical steganography uses scientific methods to hide a message, such as the use of invisible ink or microdots and other size reduction methods. Linguistic steganography hides the message within the carrier in some non-obvious ways and is further categorized as semagrams or open codes.

On computers and networks, stego applications allow someone to hide any kind of binary file into any other binary file, although today's most common carriers are the image and audio files.

Further methods of text steganography, known and developed by the authors of this article, will be analyzed from the point of view of their efficiency (or steganography capacity).

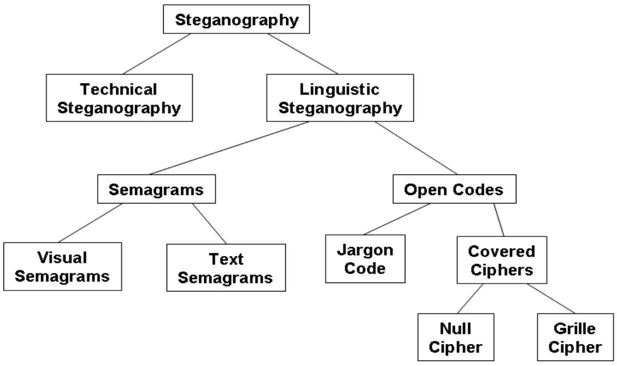


Fig. 1. The classification of steganography techniques

The characteristic of base methods of text steganography

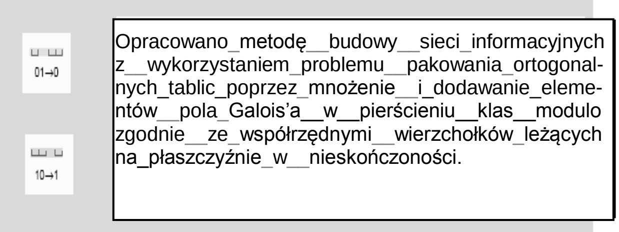
The feature of text steganography is a relative deficiency in the container of a surplus information and thereof small volume built in stego-medium. Other methods give an opportunity to bring in the image or a sound updatings imperceptible to an eye or ear while an additional letter or a sign of the punctuation in the text can be easily recognized by the casual reader.

All variety of methods of the text steganography is subdivided into syntactic and linguistic methods. Syntactic methods do not affect semantics of the text message. Linguistic methods are based on equivalent transformation of text files. These methods keep the semantic maintenance of the text untouched.

The following concerns the cores of known methods.

A text semagram hides a message by modifying the appearance of the carrier text, such as subtle changes in font size or type, adding extra spaces, or different flourishes in letters or handwritten text.

First of analyzed methods (Fig. 2) assumes accommodation of the information on the basis of change of quantity of signs "space" in each pair. The sign "space" is replaced with a sign "underlining". So, for example, the first line of the message in Fig. 2 contains the two bits of information: 01, while the second line contains another two bits - 11; the third line contains one bit -1 etc. Thus, the container contains the following message: 011111110. The important feature is that the text is formatted from both sides.



Opracowano_metodę__budowy__sieci_informacyjnych
z__wykorzystaniem_problemu__pakowania_ortogonal-
nych_tablic_poprzez_mnożenie__i_dodawanie_eleme-
ntów__pola_Galois’a__w__pierścieniu__klas__modulo
zgodnie__ze_współrzędnymi__wierzchołków_leżących
na_płaszczyźnie_w__nieskończoności.

Fig. 2. An example of the use of spaces of different length in the text-container

The second method is based on artificial increase in length of a line (Fig. 3). Here the key factor is the quantity of spaces after visible last symbol; for example, if there is no spaces, the value of the bit is *b* = 0; if there is one space, b = 1. Other principle of coding can also be applied.

Opracowano metodę budowy sieci informacyjnych z
wykorzystaniem problemu pakowania ortogonalnych tablic_
poprzez mnożenie i dodawanie elementów pola Galois’a w_
pierścieniu klas modulozgodnie z współrzędnymi wierz-
chołkó leżących n płaszczyźni w nieskończoności.

Fig. 3. An example of the use of a method based on artificial increase in length of a line

In the text in Fig. 3 the conveyed information is: 01100. Two following methods: *Line-Shift Coding* and *Word-Shift Coding*, are presented and described in detail in [5].

The Line-Shift Coding method of altering a document by vertically shifting the locations of text lines to encode the document uniquely. This encoding may be applied either to the format file or to the bitmap of a page image. The embedded codeword may be extracted from the format file or bitmap. In some cases this decoding can be accomplished without need of the original image, since the original is known to have uniform line spacing (i.e., “leading”) between adjacent lines within a paragraph.

Fig. 4 illustrates an example of use of *Line-Shift Coding* method. The last line has been shifted up by 1 pt.

Opracowano metodę budowy sieci informacyjnych z
wykorzystaniem problemu pakowania ortogonalnych tablic
poprzez mnożenie i dodawanie elementów ciała Galois’a w
pierścieniu klas modulozgodnie z współrzędnymi wierz-
chołków leżących na płaszczyźnie w nieskończoności.

Fig. 4. An example of the use of *Line-Shift Coding* method

To this example, a case of accommodation in the text the information 00001 can be adequate. Other principle of coding is also possible. Its main idea is that the text with three various interlower case distances is used. It is allocated the maximal and minimal distances between the lines designating accordingly symbol 1 and symbol 0. Other distances increase or reduce till the sizes allocated.

The Word-Shift Coding method of altering a document by horizontally shifting the locations of words within text allows to encode the document uniquely. This encoding can be applied to either the format file or to the bitmap of a page image. Decoding may be performed from the format file or bitmap. The method is least visible when applied to documents with variable spacing between adjacent words. Variable spacing in text documents is commonly used to distribute white space when justifying text.

Because of this variable spacing, decoding requires the original image - or more specifically, the spacing between words in the unencoded document. See Fig. 5 for an example of word-shift coding.

a)

Opracowano metodę budowy sieci informacyjnych
Opracowano metodę budowy sieci informacyjnych

b)

Opracowano metodę budowy sieci informacyjnych
Opracowano metodę budowy sieci informacyjnych

Fig.5. An example of Word-Shift Coding

In Fig. 5a the top text line has added spacing (0.1 mm) before the "metodę" (in the first line) and the same spacing before the "budowy" (in the second line); the bottom text line has the same spacing after the "budowy". In Fig. 5b the same text lines without the vertical lines are shown to demonstrate that either spacing looks natural.

The next method is the most effective, however it is the most trivial and clear one (see Fig. 6): let a symbol of the bottom register, for example, to be corresponding to 0, and a symbol of the top register – to 1. Thus in last figure the information 0100000001000 is besieged.

«sTeganogrAphy»

Fig. 6. A simple illustration of the usage of register

For an estimation of "capacity" of the container for some methods, their comparative analysis on a given example is performed. As the container the same text file is used and the same message is introduced in it by different methods. Such a message was the word "*Здравствуйте*" (in Russian).

The results of the performed experiment for three simple methods are shown in Tab. 1.

Table 1. Results of the executed experiment

Method	File size, in bytes	The number of signs and spaces
Container	34304	2992
Spaces of different length	36864	3090
Method based on artificial increase in the length of line	34304	3007

Feature Coding. This method of coding is applied either to a format file or to a bitmap image of a document. The image is examined for chosen text features, and those features are altered, or not altered, adequately to the codeword. The decoding requires the original image, or more precisely, a specification of the change in pixels in a feature. There are many possible choices of text features; exemplarily, the upward, vertical endlines - that is the tops of letters *b*, *d*, *h*, etc. can be altered. These endlines are altered by extending or shortening their lengths by one (or more) pixels, but otherwise not changing the endline feature [5]. It is obvious, that the given method is hardly sold.

The method of an identical tracing of symbols is based on the fact that the display of the majority of standard fonts (Times New Roman, Arial, Courier New and others) of different symbols of both the Russian and the English alphabet on the screen is identical. So, the Russian symbols "уехарсосЕХАРОСНМТВ" and the English symbols

-"уехаросEXAPOCHMTB” cannot be distinguished from each other at reading and viewing of the text. Thus, embedding stego (hidden information) in the container is possible by replacement of a symbol of the Russian alphabet by the same, identically displayed symbol of the English alphabet, accepting as 1 the letter of Russian allocation of the keyboard, and as 0 – the English allocxation, or vice versa. After carrying out the embedding of the stego in the container, at any change it will not be appreciable. But thus, the great part of the filling of the container will be provided by data, as the number of replaced symbols is large enough: 17 from 52 (symbols of the top and bottom register of the English alphabet) and 17 from 66 (symbols of the top and bottom register of the Russian alphabet). And these symbols possess high frequency of repeatability in the text of any maintenance.

The efficiency of the given method of packing hidden information in the container has been investigated on the text of the first volume of the L.N.Tolstoy novel "War and peace" in volume of 1 472 844 bytes, translated in an ASCII-code, with quantity of symbols 1 460 024. In the given source it is possible to hide 484 907 bytes of useful information, that means 33,2 % from total of symbols in the container.

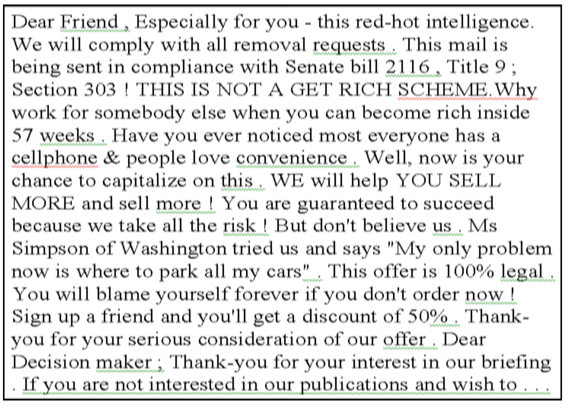
The Open Codes hide a message within a legitimate carrier message in ways that are not obvious to an unsuspecting observer. The carrier message is sometimes called “the overt communication” while the hidden message is “the covert communication”. This category is subdivided into *jargon codes* and *covered ciphers*.

Jargon code, as the name suggests, uses language that is understood by a group of people but is meaningless to others. Jargon codes include warchalking (symbols used to indicate the presence and type of wireless network signal [6]), underground terminology, or an innocent conversation that conveys special meaning because of facts known only to the speakers. A subset of jargon codes are *cue codes*, where certain pre-arranged phrases convey meaning.

Covered, or concealment, ciphers hide a message openly in the carrier medium so that it can be recovered by anyone who knows the secret how it was concealed. A *grillecipher* employs a template that is used to cover the carrier message; the words that appear in the openings of the template are the hidden message. A *null cipher* hides the message according to some prearranged set of rules, such as "read every fifth word" or "look at the third character in every word".

One of the simplest null ciphers is shown in the example below. Let the channel the message is sent be open: `Ala ma firmę, oferującą rejestratory alarmu`. Reading the first character of every word in the message will yield the hidden text: `Amfora`.

The second method of this group is called *Spammimic* because the container adopts an usual spam (or any neutral text) inside which the meaning symbols (stego) are placed in an established way. Fig. 7 represents an example of such message. This message looks like the spam that most of us receive every day, which we ignore and discard. This message was created at spammimic, a Web-site that converts a short text message into a text block that looks like spam using a grammar-based mimicry idea, first proposed in [7]. The reader will learn nothing by looking at the word spacing or misspellings in the message; the zeroes and ones are encoded by the very choice of the words. The hidden message in the "spam carrier" below is: `Здравствуйте`.



Dear Friend , Especially for you - this red-hot intelligence. We will comply with all removal requests . This mail is being sent in compliance with Senate bill 2116 , Title 9 ; Section 303 ! THIS IS NOT A GET RICH SCHEME.Why work for somebody else when you can become rich inside 57 weeks . Have you ever noticed most everyone has a cellphone & people love convenience . Well, now is your chance to capitalize on this . WE will help YOU SELL MORE and sell more ! You are guaranteed to succeed because we take all the risk ! But don't believe us . Ms Simpson of Washington tried us and says "My only problem now is where to park all my cars" . This offer is 100% legal . You will blame yourself forever if you don't order now ! Sign up a friend and you'll get a discount of 50% . Thank-you for your serious consideration of our offer . Dear Decision maker ; Thank-you for your interest in our briefing . If you are not interested in our publications and wish to . . .

Fig. 7 An example of spam-message with hidden information

The Method “Spammimic” was investigated by the authors on the basis of the software product accessible on a site [8]. As the container the same file was used, specified in Tab. 1. Thus, it is revealed that the size of a file with the “built in” information is 26624 bytes, while the number of signs and blanks is 1890. It is less than at realization of methods from Tab. 1.

Conclusion

A set of experiments and calculations was conducted by the authors to demonstrate that small variations in line spacing ,indiscernible to a casual reader, can be recovered from a paper copy of the document, even after being copied several times.

In such methods as *Feature Coding* or the increase in the register of letters, the concealment of the message is most obvious.

The method externally most imperceptible is the Null cipher method.

REFERENCES

[1] Eason R., Digital Steganography, Proc. of Pacific Rim Workshop on Digital Steganography, 2002, 1-6

[2] Kopniak P., Zabezpieczenie informacji poprzez jej ukrycie-steganografia i jej narzędzia, W: Bezpieczeństwo informacji: od teorii do praktyki, Marek Miłosz, Warszawa, MIKOM, 2005, 145-156

[3] Chandramouli R., Data hiding capacity in the presence of an imperfectly known channel, SPIE Proceedingsof Security and Watermarking of Multimeida Contents II 4314, 2001

[4] Bauer F.L., Decrypted Secrets: Methods and Maxims of Cryptology, 3rd ed. New York:Springer-Verlag, 2002

[5] Brassil J.T., Low S., Maxemchuk N.F., O’Gorman L., Electronic Marking and Identification Techniques to Discourage Document Copying, *IEEE Journal On Selected Areas In Communications*, 13 No.8 (1995), 1495-1504

[6] Warchalking Web Site. Warchalking: Collaboratively creating a hobo-language for free wireless networking, URL: http://www.warchalking.org: 2003-12-21

[7] Wayner P., Disappearing Cryptography - Information Hiding: Steganography &Watermarking, 2nd. ed. San Francisco: Morgan Kaufmann, 2002

[8] Spammimic, Web Site. URL: http://www.spammimic.com: 2009-03-29

Authors:

prof. dr hab. inż. Paweł Urbanowicz, studenci Konstantin Czurikow, Andrej Rimorew, Nadzieja Urbanowicz, Białoruski Państwowy Uniwersytet Technologiczny, ul. Swerdlowa 14a, 220000 Mińsk; Katolicki Uniwersytet Lubelski, al. Racławickie, 14, Lublin, E-mail: upp@rambler.ru; nadya_ur@rambler.ru.;