A New Synonym Text Steganography

M. Hassan Shirali-Shahreza
Computer Engineering Department
Yazd University
Yazd, IRAN
hshirali@yazduni.ac.ir

Mohammad Shirali-Shahreza
Computer Science Department
Sharif University of Technology
Tehran, IRAN
shirali@cs.sharif.edu

Abstract

Steganography is a relatively new method for establishing hidden communication which gained attraction in recent years. Steganography is a method of hiding a secret message in a cover media such as image or text. In this paper a new method is proposed for steganography in English text by substituting the words which have different terms in British English and American English.

1. Introduction

Internet has growth rapidly in recent years. One of the areas which is gained attracted by many people is security subjects on the Internet. Among these topics, nowadays establishing hidden communication is a hot topic that received more attentions.

Various methods including cryptography, steganography, coding, etc. are used for establishing hidden communication. Steganography is the process of hiding data inside other data in such a way that no one apart from the intended recipient knows of the existence of the message. This is the major distinction between steganography and other methods of hidden exchange of information. For example, in cryptography method, people become aware of the existence of information by observing coded information, although they are unable to comprehend the information.

Most steganography jobs have been performed on images, video clips, text, music and sound [1]. But text steganography is the most difficult kind of steganography; this is due to the lack of redundant information in a text file, while there is a lot of redundancy in a picture or a sound file, which can be used in steganography [2].

The structure of text documents is identical with what we observe, while in other types of documents such as in picture, the structure of document is different from what we observe. Therefore, in such documents, we can hide information by making changes in the structure of the document without making a notable change in the concerned output.

Accordingly, a few works have been done on hiding information in texts. The survey on some of these methods is available in next section.

In this paper a new method for steganography in English text is presented. In this method the words which have different terms in UK and US are substituted in order to hide data in an English text. In English some words have different term in US and UK. For example “lift” has different terms in UK (elevator) and US (lift). So we can hide data in the text by substituting these words. The details of this method are explained in third section.

Among the text steganography methods reported so far, Semantic Method [3] is similar to our method. This method is explained in next section.

In the final section the conclusion will be made after investigating and studying some advantages of this method.

2. Related works

As we said in previous section, a few works have been done on hiding information in texts. Here, we make a review of some works done on the text steganography.

2.1. The line shifting [4]

In this method, the lines of the text are vertically shifted to some degree (for example, each line is shifted 1/300 inch up or down) and information are

978-0-7695-3278-3/08 $25.00 © 2008 IEEE
DOI 10.1109/IIH-MSP.2008.6

hidden by creating a unique shape of the text. This method is suitable for printed texts.

However, in this method, the distances can be observed by using special instruments of distance assessment and necessary changes can be introduced to destroy the hidden information. Also if the text is retyped or if character recognition programs (OCR) are used, the hidden information would get destroyed.

2.2. Word shifting [5]

In this method, by shifting words horizontally and by changing distance between words, information is hidden in the text. This method is acceptable for texts where the distance between words is varying. This method can be identified less, because change of distance between words to fill a line is quite common.

But if somebody was aware of the algorithm of distances, he can compare the present text with the algorithm and extract the hidden information by using the difference. The text image can be also closely studied to identify the changed distances. Although this method is very time consuming, there is a high probability of finding information hidden in the text. The same as in the method described under 2.1, retyping of the text or using OCR programs destroys the hidden information.

2.3. Syntactic method [6]

By placing some punctuation marks such as full stop (.) and comma (,) in proper places, one can hide information in a text file.

This method requires identifying proper places for putting punctuation marks. The amount of information to hide in this method is trivial.

2.4. Semantic method [3]

This method uses the synonym of certain words thereby hiding information in the text. A major advantage of this method is the protection of information in case of retyping or using OCR programs (contrary to methods listed under 2.1 and 2.2).

However, this method may alter the meaning of the text.

2.5. Text abbreviation [2]

Another method for hiding information is the use of abbreviations.

In this method, very little information can be hidden in the text. For example, only a few bits of information can be hidden in a file of several kilobytes.

There are also other text steganography methods which are not similar to our method such as Open Spaces [7] and Feature Coding [8].

3. Our suggested method

The goal of our method is to hide data in a text by substituting the words which have different terms in UK and US. For example “Pants” has different terms in UK (Trousers) and US (Pants). So we can hide data in the text by substituting these words. Table I shows a number of such words which have different terms in UK and US. The details of implementing this method are as follows:

We have used Java programming language to implement this method. This project is composed of two programs:

1- Hiding program which is responsible for hiding data in text.

2- Extractor program which extracts data from the stego text (text containing hidden data).

At first we prepare a list or dictionary containing the words which have different terms in UK and US.

The hiding program looks for existing words in the list in the text. Furthermore this program converts the concerned data to an arrangement of 0 and 1 bits. The program will place US term in sentence for hiding of the bit 0 and will place the UK term in the sentence in order to hide the bit 1.

This way the data will be hidden in the concerned text. Of course the size of data is hidden in the text in order that the extractor program can work correctly.

The extractor program extracts the data from the stego text. This program identifies the type of words in stego text by using the list of words having different terms in UK and US and saves the quantity of 0 or 1 in an arrangement according to the fact whether it is a US term or a UK term. Now the hidden data will be extracted through conversion of this arrangement from the bits 0 and 1 to its original format. At the end the extracted data will be saved on the user's computer.

Table 1. List of some words which have different terms in UK and US

American English	British English
Account	Bill
Candy	Sweets
Closet	Cupboard
Faculty	Staff
Fall	Autumn
Gas	Petrol
Incorporated	Limited
Mail	Post
Movie	Film
Package	Parcel
Soccer	Football
Stove	Cooker

This method has little capacity to hide data in text. However this is related to the body of text and its size, but in overall its capacity is very low. Although its capacity is low, but we must be note the situation that this method is used, because in some situations we need to hide little data, but other steganography methods are not suitable and can be broken. Also our suggested method is a new one, therefore the possibility of breaking this method is low. On the other hand, in countries which their native language is not English and their English language knowledge is low, substituting US and UK words are not attract attention much, so this method is applicable in these countries.

4. Conclusion

This paper presents a new text steganography method for English texts.

In English some words have different term in US and UK. For example “rubber” has different term in UK (eraser) and US (rubber). By using this feature, we propose our method for hiding data in an English text. In this method we hide data in the text by substituting such words.

This method is not only for electronic documents and can be used on printing texts. Also by printing the electronic document, the hidden data is not destroyed and will remain.

In some languages, people of different cities have their local accent. Therefore there are also some words in their languages which have different terms. For example in Persian, the word “Divar” (meaning Wall) has different term in some cities and said “Chine”. So we can also use this feature in our method.

Our suggested method is new, so the possibility of breaking this method is low. Cryptography of the intended data can also add the security of this method.

5. References

[1] N.J. Hopper, *Toward a theory of Steganography*, Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, July 2004.

[2] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding", *IBM Systems Journal*, vol. 35, issues 3&4, 1996, pp. 313-336.

[3] M. Niimi, S. Minewaki, H. Noda, and E. Kawaguchi, "A Framework of Text-based Steganography Using SD-Form Semantics Model", *Pacific Rim Workshop on Digital Steganography 2003*, Kyushu Institute of Technology, Kitakyushu, Japan, 3-4 July 2003.

[4] A.M. Alattar and O.M. Alattar, "Watermarking electronic text documents containing justified paragraphs and irregular line spacing", *Proceedings of SPIE -- Volume 5306, Security, Steganography, and Watermarking of Multimedia Contents VI*, June 2004, pp. 685-695.

[5] Y. Kim, K. Moon, and I. Oh, "A Text Watermarking Algorithm based on Word Classification and Inter-word Space Statistics", *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR’03)*, 2003, pp. 775–779.

[6] K. Bennett, "Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text", *Purdue University*, CERIAS Tech. Report 2004-13, 2004.

[7] D. Huang and H. Yan, "Interword Distance Changes Represented by Sine Waves for Watermarking Text Images", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 12, December 2001, pp. 1237-1245

[8] K. Rabah, "Steganography-The Art of Hiding Data", *Information Technology Journal*, vol. 3, issue 3, 2004, pp. 245-269.