Text Steganography by Changing Words Spelling

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

Abstract — **One of the important issues in security fields is hidden exchange of information. There are different methods for this purpose such as cryptography and steganography. Steganography is a method of hiding data within a cover media so that other individuals fail to realize their existence. In this paper a new method for steganography in English texts is proposed. In this method the US and UK spellings of words substituted in order to hide data in an English text. For example "color" has different spelling in UK (colour) and US (color). Therefore the data can be hidden in the text by substituting these words.**

Keywords — **Data Hiding, Information Security, Text Steganography, Word Spelling.**

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 receives 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 will be 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].

Accordingly, a few works have been done on hiding information in texts, such as Word Shifting, Syntactic Methods, Semantic Methods, Open Spaces, etc. In the next section some of these methods are reviewed.

In this paper a new method for steganography in English text is presented. In this method the US and UK spellings of words substituted in order to hide data in an English text. For example "analyze" has different spelling in UK (analyse) and US (analyze). The details of this method are described in the third section.

This method is implemented by Java programming language. In the third section we will discussed how to implement the project.

Among the reported text steganography methods, Semantic Methods [2] is similar to our method. In the next section this method is explained.

In the final section the conclusion will be made after 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 methods.

2.1. Line Shifting [4]: In this method, the lines of the text are vertically shifted to some degree (for example, each line shifts 1/300 inch up or down) and information are hidden by creating a unique shape of the text. This method is proper 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 are 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 previous method, retyping of the text or using OCR programs destroys the hidden information.

2.3. Syntactic Methods [6]: By placing some punctuation signs 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 signs. The amount of information to hide in this method is trivial.

2.4. Semantic Methods [2]: This method is similar to our method. In this method, the synonym of words is used for certain words thereby the information is hidden 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. 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 can be hidden in a file of several kilobytes.

Also there are other text steganography methods such as Feature Coding and Open Spaces. The survey on these methods is available in [7].

3. The Suggested Method

This project offers a new steganography method for hiding data in English texts. In English some words have different spelling in UK and US. For example "dialog" has different terms in UK (dialogue) and US (dialog). So we can hide data in the text by substituting these words. Table I shows a number of such words which have different spelling in UK and US. The details of this method and its implementation 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 text containing hidden data.

At first we prepare a list containing the words which have different UK and US spelling.

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 word in sentence for hiding of the bit 0 and will place the UK word 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 will extract the data from the text. This program identifies the type of words in text by using the list of words having different UK and US spelling and saves the quantity of 0 or 1 in an arrangement according to the fact whether it is a US word or UK word. 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.

This method has little capacity to hide data in the 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, 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.

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

American Spelling	British Spelling
Favorite	Favourite
Criticize	Criticise
Fulfill	Fulfil
Center	Centre
Dialog	Dialogue
Medieval	Mediaeval
Check	Cheque
Defense	Defence
Tire	Tyre

On the other hand, in countries which their native language is not English and their English language knowledge is low, substituting US and UK spellings of 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 hiding data in English texts. This method is based on substituting US and UK spellings of words.

In English some words have different spelling in US and UK. For example "program" has different spelling in UK (programme) and US (program). 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. By printing the electronic document, the hidden data is not destroyed and will remain.

Our suggested method is new, so the possibility of breaking this method is low.

This method can be used on other devices such as Pocket PC, PDAs and mobile phones. For example this method can be applied on SMS (Short Message Service) messages. The SMS is a popular service on mobile phones which is the transfer and exchange of short text messages between mobile phones.

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.

[7] M.H. Shirali-Shahreza and M. Shirali-Shahreza, "Text Steganography in Chat", Proceedings of the Third IEEE/IFIP International Conference in Central Asia on Internet the Next Generation of Mobile, Wireless and Optical Communications Networks (ICI 2007), Tashkent, Uzbekistan, September 26-28, 2007.