Journal of Theoretical and Applied Information Technology

www.jatit.org

STEGANOGRAPHY IN PERSIAN AND ARABIC UNICODE TEXTS USING PSEUDO-SPACE AND PSEUDO CONNECTION CHARACTERS

[1]M. Hassan Shirali-Shahreza, [2]Mohammad Shirali-Shahreza

[1]Assistant Prof., Computer Engineering Department, Yazd University, Yazd, IRAN

[2]Student, Computer Science Department, Sharif University of Technology, Tehran, IRAN

E-mail: shahreza@shirali.ir, shirali@cs.sharif.edu

ABSTRACT

Sending information secretly and communicating covertly have been of great interest for ages. On the other hand, text documents have been widely used and consequently various methods for hiding information in texts (Text Steganography) have been developed so far. In this paper a new method is proposed for hiding information in Persian and Arabic Unicode texts. In Persian and Arabic, some letters are connected together in a word. But there are two characters, zero width non joiner (ZWNJ) and zero width joiner (ZWJ) characters, which are respectively prevents the Persian and Arabic letters from joining or forces them to join together. In this method by using these two special characters, the information is hidden in Persian and Arabic Unicode text documents. This method has a high hiding capacity because it hides one bit in each letter. Also this method does not make any apparent changes in the original text and have a perfect perceptual transparency.

Keywords: *Text Steganography, Persian/Arabic Text, Pseudo-Connection, Pseudo-Space, Unicode Standard.*

1. INTRODUCTION

In the 21st century, communications are expanded because of developing new technologies such as computers, the Internet, mobile phones, etc. By using these technologies in different areas of life and work, the issue of information security has gained special significance. Hidden exchange of information is one of the important areas of information security which includes various methods such as cryptography, Steganography, and coding.

In Steganography the information is hidden in a cover media so nobody notice the existence of the secret information. Steganography works have been carried out on different medium such as images, videos and sounds [1].

Text Steganography is the most difficult kind of Steganography because there is a little redundant information in a text file as compared with a picture or a sound file [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 introducing changes in the structure of the document without making a notable change in the concerned output.

Contrary to other media such as sounds and video clips, using text documents has been common since very old times. This has extended until today and still, using text is preferred over other media, because the texts occupy lesser space, communicate more information and need less cost for printing as well as some other advantages.

As the use of text and hidden communication goes back to antiquity, we have witnessed to Steganography in texts since past. For example, this method has been done by some Iranian classic poets as well.

Today, the computer systems have facilitated information hiding in texts. The applications of information hiding in text have also expanded from hiding information in electronic texts and documents to hide information in web pages [3].

Most of the text Steganography methods are for English texts and there are a few text Steganography methods for other languages.

In this paper we introduce a new text steganography method for Persian and Arabic texts saved in Unicode format. Our method is based on zero width non joiner (ZWNJ) and zero width joiner (ZWJ) characters which are also known as Pseudo-Space and Pseudo-Connection.

The details of our method are described in the third section. In section four, we will discuss about the experimental results of this method and analyze it.

As we said earlier, there are a few text Steganography methods for non English languages. In the next section, we will discuss some reported Persian and Arabic text Steganography methods.

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

2. RELATED WORKS

In this section, the text Steganography methods that are especially designed for Persian and Arabic texts are surveyed.

To the best knowledge of the authors, there are only four Persian and Arabic text Steganography methods that are reported in the literatures. The first two methods were developed by the authors of this article.

In the Dot Steganography method [4], data is hidden in Persian and Arabic texts by using a special characteristic of these languages. Considering the existence of too many dots in Persian and Arabic characters, in this approach by vertical displacement of the dots (see Figure 1), we hide information in the texts. This method does not attract attention and can hide a large volume of information in text.

The La Steganography method [5] uses the special form of "La" word for hiding the data. This word is created by connecting "Lam" and "Alef" characters. For hiding bit 0, we use the normal form of word "La" ("لا") by inserting Arabic extension character between "Lam" and "Alef" characters. But for hiding bit 1, we use the special form of word "La" ("ﻻ") which has a unique code in the Unicode Standard (its code is FEFB in Unicode hex notation). This method is not limited to electronic documents (e-documents) and can also be used on printed documents.

Figure 1. Vertical displacement of the dots of the Persian letter NOON [4]

The third method [6], uses the pointed letters with extension (Kashida in Arabic) to hold secret bit 'one' and the un-pointed letters with extension to hold secret bit 'zero'. Note that letter extension does not have any effect to the writing content. It has a standard character hexadecimal code of 0640 in the Unicode system.

The extension is added before (or after) the pointed letters which can be extended with extension character to hide bit 1, and added before (or after) the un-pointed letters to hide bit 0. Table 1 shows an example of this method.

In the fourth method [7], a diacritic Arabic text is used for hidden exchange of information. There are eight diacritics in Arabic text. The most frequent diacritics in Arabic text is "Fatha" and the probability of its occurrence is equal to the occurrence probability of other seven diacritics [7].

In this method, at the first the cover text is assumed to be a fully diacritical text. To hide a bit "1" a "Fatha" is kept and to hide a bit "0" a non "Fatha" diacritic is kept and other diacritics are removed. So in the stego text each "Fatha" represents "1" and each non "Fatha" diacritic represents a "0".

The main advantage of this method is its high capacity. But the main disadvantage of this method is that it attracts the attention of the reader. This method also needs a fully diacritical text, but most of Arabic texts have no diacritic.

Table 1. Example of text Steganography method proposed in [6] (adding extensions after letters)

Watermarking bits	110010
Cover-text	من حسن اسلام المرء تركه مالا يعنيه
Output text	مـن حسن اسلام المرء تـركـه مـالا يـعنـيه 1 1 0 0 1 0

3. OUR PROPOSED METHOD

In this paper, we present a new method for text Steganography in Persian and Arabic Unicode texts.

Before explaining the method, we mention the main characteristics of these two languages [8]. Then we explain the Unicode Standard briefly and at last we explain our suggested method in full details.

3.1 THE CHARACTERISTICS OF PERSIAN AND ARABIC

Arabic alphabet has 28 letters. Persian has all the letters of Arabic and four more letters of (پ،چ،ژ،گ). In these two languages, a letter can have four different shapes. The shape of each letter is determined by the position of that letter in a word. For example the letter «ع» is written as «عـ» at the beginning of a word, as «ـعـ» in the middle, as «ـع» at the end, and as «ع» in the isolated form.

In Persian and Arabic the letters are connected to each other in both handwritten and printed texts, while in the English, the letters are written separately in printed texts.

In English, the letters are written in a left-to-right format and in some languages the letters are written in a top-to-bottom format, but in Arabic and Persian the letters are written in a right-to-left format.

In Arabic and Persian languages, dot is very important and 17 of 32 Persian letters (and 14 of 28 Arabic letters) have one or more dots. Among these 17 letters, 2 letters have 2 dots and 5 letters have 3 dots and the remaining 10 letters have one single dot, while in English only two small letters "i" and "j" have dot.

In Persian and Arabic some letters do not connected to each other. The zero width joiner (ZWJ) is a non-printing character which is when placed between two characters that would otherwise not be connected, a ZWJ causes them to be printed in their connected forms. It is also known as pseudo-connection. The ZWJ's Unicode is U+200D.

But also there is a non-printing character in Persian and Arabic which prevents the Persian and Arabic letters from joining without adding a space between the two and keeps the words closer together. This character is named zero-width non-joiner (ZWNJ). It is also known as pseudo-space or non-breakable space (NBSP). Its code is U+200C in Unicode hex notation.

We use both of these characters of Arabic and Persian languages in our method and will explain it in this section.

3.2 UNICODE STANDARD

Unicode Standard [9] is the international character-encoding standard used for presenting the texts to process by computers. This standard is compatible to the second version of ISO/IEC 10646-1:2000 and has the same characters and codes of ISO/IEC 10646.

The Unicode standard enables us to encode all the characters used in writing of the world languages. This standard uses the 16-bit encoding which provides space for 65000 characters. So it is possible to specify and define 65000 characters in different moulds such as numbers, letters, symbols, and a great number of current characters in different languages of the world.

The Unicode standard has determined codes for all the characters used in main languages of the world. Moreover, because of the wideness of the space dedicated to the characters, this standard also includes most of the symbols necessary for high-quality typesetting. The languages whose writing system can be supported by this standard are Latin (covering most of the European languages), Cyrillic (Russian and Serbian), Greek, Arabic (including Arabic, Persian, Urdu, Kurdish), Hebrew, Indian, Armenian, Assyrian, Chinese, Katakana, Hiragana (Japanese) and Hangeul (Korean).

Moreover there are a lot of mathematical and technical symbols, punctuation marks, arrows, and miscellaneous marks in this standard.

In the Unicode standard, the Persian characters belong to the Arabic block. This block has been developed to cover the characters of the languages which use Arabic writing system. Among these languages we can mention Persian, Urdu, Pashto, Sindhi, and Kurdish.

This standard has detailed and careful explanations about the implementation methods including letters-connection method, the exhibition of the right-to-left and bi-direction texts. This way the programmers do not have to refer to the local guide.

3.3 OUR METHOD

In this paper we introduce a new text steganography method for Persian and Arabic texts. Our method is based on zero width non joiner (ZWNJ) and zero width joiner (ZWJ) characters

which are also known as Pseudo-Space and Pseudo-Connection..

As we said in section 3.1, in Persian and Arabic, some letters are connected together in a word, but other letters cannot join together. ZWNJ prevents the Persian and Arabic letters from joining, but ZWJ forces the letters to join together.

The method proposed in this paper for hiding data in Persian and Arabic Unicode texts is using both of these characters.

In this method we hide one bit in each letter. For hiding data in this method, first we look whether the letter of a word is connected to the next letter or not. If it is connected to the next letter, we insert ZWJ letter between two letters for hiding bit 1 and do not add anything for hiding bit 0. Because the letters are connected together, adding ZWJ for connecting the letters together does not have any effects on the apparent of the text.

But if the letter is not connected to the next letter, we insert ZWNJ letter between two letters for hiding bit 1 and do not add anything for hiding bit 0. Also in this case the apparent of the word is not changed; because the letters are not connected together and adding ZWNJ for separating the letters from each other does not have any effects on the apparent of the word. For hiding data in the last letter of a word, we always insert ZWNJ letter after it for hiding bit 1 and do not add anything for hiding bit 0.

In Persian and Arabic, in addition to the space that is provided between words, in some words such as "می‌باشد", there is a small space between the two parts of the same word, i.e. "می" and "باشد". For inserting this type of space, ZWNJ letters is inserted between the two letters instead of the normal space. If we have a word of this type in our text, we insert two more ZWNJ letter between two letters for hiding bit 1 and insert one more ZWNJ letter between two letters for hiding bit 1. Therefore the hidden data can be detected correctly in extracting phase.

To extract the information from the text having hidden information (stego text), we respectively investigate the letters of the text words. If after the letter there is a one or three ZWNJ or one ZWJ character, it means that the bit 1 is hidden in that word. But if after the letter there is no ZWNJ and ZWJ or there is two ZWNJ, it means that the bit 0 is hidden in this letter. By putting all the bits of 0 and 1 next to each other we can extract the hidden information from the text.

Because ZWJ and ZWNJ character are a non-printing characters, therefore this method does not make any apparent changes in the apparent original text and have a perfect perceptual transparency.

Also our method has very high hiding capacity, because we hide one bit in each letter.

4. EXPERIMENTAL RESULT

In our method, the information is hidden in Persian and Arabic Unicode texts using both ZWNJ and ZWJ characters.

We tested our method on some Persian text files. We selected the resources which are used in our previous Persian and Arabic text Steganography methods [3, 4] in order to compare these methods.

These resources are selected for computing the capacity of the methods for hiding data and including sport pages of some Iranian newspapers. The Internet address of these newspapers and the capacity of each text for hiding data are shown in Table 2. All of the pages were retrieved on 20 August 2005.

As it is seen in the table 2, our method capacity is very high, especially in comparison with La Steganography method [5]. In this method we hide a bit of information in each Persian and Arabic letter. So our method capacity is four times higher than Dot Steganography method [4] in average.

Also our method has advantages over these methods. For example, contrary to the Dot Steganography method [4], this method does not change the apparent of the text and does not required specific font. We will discuss some advantages of our method in the next section.

5. CONCLUSION

In this paper, a new method for Steganography in Persian and Arabic Unicode texts has been presented. In this method for hiding bit 1 Pseudo-Space or Pseudo-Connection characters is inserted between two letters regarding to whether they are connected or not, for hiding bit 0 no change is made.

This method is not dependent on any special format and we can save the stego text in numerous formats such as HTML pages, Microsoft Word documents or even plain text format. Because the stego Unicode texts will not change during copy and paste between computer programs, the data hidden in texts remains intact during these operations.

Table 2. Example of text Steganography method proposed in [6] (adding extensions after letters)

Newspaper	WebSite Address	Text Size (Kilo Byte)	Our Method Text Capacity (bit)	Our Method Capacity Ratio (Bit/ Kilobyte)	Dot Steganography Method [4] Capacity Ratio (Bit/ Kilobyte)	La Steganography Method [5] Capacity Ratio (Bit/ Kilobyte)
Farhange Ashti	http://www.ashtidaily.com	13.3	5640	424	96	1.43
Hamshahri	http://www.hamshahri.net	6.82	2770	406	120	1.03
Iran	http://www.iraninstitute.org	6.64	2707	408	105	1.66
JameJam	http://www.jamejamdaily.net	3.84	1556	405	113	1.48
Javan	http://www.javandaily.com	8.03	3226	402	115	1.00
Jomhouri Eslami	http://www.jomhourieslami.com	3.52	1413	401	125	1.14
Keyhan	http://www.kayhannews.ir	2.92	1181	404	106	2.05
Khorasan	http://www.khorasannews.com	5.40	2213	410	116	0.74
Quds	http://www.qudsdaily.net	9.98	4044	405	114	0.30
Shargh	http://www.sharghnewspaper.com	20.4	8307	407	118	0.88

There are three important parameters in designing Steganography methods: perceptual transparency, robustness and hiding capacity. These requirements are known as “the magic triangle” and are contradictory [10].

Our method satisfies both perceptual transparency and hiding capacity requirements. We did not make any apparent changes in the original text by hiding data. So even if the reader has the original text, it is impossible for him to realize the hiding of the data by merely observing the appearance of the text. However, the original texts are not available to the observer in text Steganography methods usually. Therefore the main goal of text Steganography, that is the impossibility of detection of the presence of data, has been achieved. Also we hide one bit per letter in the text file, so our method has ultra high capacity is high.

In some Steganography methods, standard structure of the text will be disarranged and spelling and grammatical errors will be created in the text, but in this method the appearance of the text will not change at all and the text still remains standard.

The Unicode standard supports different languages and can be used on different systems and devices which are supporting the Unicode Standard. Moreover, the Arabic is the official language of the Muslims and about two billion Muslims live throughout the world. As a result, a wide range of the users can use our method.

Since Pashto (the official language of Afghanistan) and Urdu (the official language of Pakistan) are similar to Arabic and Persian, we can also apply this method to these two languages.

In addition to the mentioned items in part 3.1, the Arabic and Persian languages have other specific characteristics which we can be used for text Steganography.

This method can be used for secret communication and for the prevention of the illegal reproduction and distribution of the texts, especially e-documents as well.

REFRENCES

[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, pp. 313-336, 1996.

[3] M. Shirali-Shahreza, "A New Method for Steganography in HTML Files", *Proceedings of the International Joint Conference on Computer, Information, and Systems Sciences, and Engineering (CISSE 2005)*, Bridgeport, CT, December 10- 20, 2005, pp. 247-251.

[4] M.H. Shirali-Shahreza and M. Shirali-Shahreza, "A New Approach to Persian/Arabic Text Steganography", *Proceedings of the 5th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2006)*, Honolulu, HI, USA, July 10-12 2006, pp. 310-315.

[5] M. Shirali-Shahreza, "A New Persian/Arabic Text Steganography Using "La" Word", *Proceedings of the International Joint Conference on Computer, Information, and Systems Sciences, and Engineering (CISSE 2007)*, Bridgeport, CT, USA, 2007.

[6] A. Gutub and M. Fattani, "A Novel Arabic Text Steganography Method Using Letter Points and Extensions", *Proceedings of the WASET International Conference on Computer, Information and Systems Science and Engineering (ICCISSE)*, Vienna, Austria, Vol. 21, 2007, pp. 28-31.

[7] M. A. Aabed, S. M. Awaideh, A. M. Elshafei, and A. A. Gutub, "Arabic Diacritics Based Steganography", *Proceedings of the International Conference on Signal Processing and Communications (ICSPC 2007)*, Dubai, UAE, 2007, pp. 756-759.

[8] M.H. Shirali-Shahreza and M. Shirali-Shahreza, "Persian/Arabic Baffletext CAPTCHA", *Journal of Universal Computer Science (J.UCS)*, Vol. 12, No. 12, December 2006, pp. 1783-1796.

[9] The Unicode Consortium, *The Unicode* Standard, http://www.unicode.org.

[10] N. Cvejic, *Algorithms for Audio Watermarking and Steganography*, Oulu University Press, Finland, 2004