Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 8 (22): 4173-4179, 2008
ISSN 1812-5654

High Capacity Persian/Arabic Text Steganography

[1]M. Shirali-Shahreza and [2]S. Shirali-Shahreza
[1]Department of Computer Science, Sharif University of Technology,
Azadi Street, P.O. Box 11365-9415, Tehran, Iran
[2]Department of Computer Engineering, Sharif University of Technology,
Azadi Street, P.O. Box 11365-9517, Tehran, Iran

Abstract: One of the methods introduced for establishing hidden communication is steganography. Steganography is the art of hiding information in a cover media without attracting attention. Text documents are one of the common cover media used for steganography in past. Steganography in text is more difficult than other media because there is a little redundant information in text documents. In this study, we propose a high capacity text steganography method. Our method can hide about 400 bits in each kilobyte of text. In Persian and Arabic, each letter can have four different shapes regarding to its position in the word. In this method this feature of Persian and Arabic languages is used for information hiding. In the Unicode Standard, there are different codes for different forms of each letter in addition to the code for letter presentation. 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 has a perfect perceptual transparency.

Key words: High capacity, perceptual transparency, Persian/Arabic text, text steganography, unicode standard

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 like cryptography, steganography and coding.

In steganography, the main objective is to hide the information in cover media so that nobody notices the existence of the secret information. 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. However, in steganography, nobody will understand the existence of information in the resources.

Steganography works have been carried out on different medium such as images, videos and sounds (Hopper, 2004). 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 (Bender *et al*., 1996).

Of course today, the security of information has been considerably improved by combination of steganography with other methods mentioned. In addition to hidden exchange of information, steganography is used in other areas such as copyright protection, preventing e-document forging and other applications (Maxemchukand and Low, 1997).

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.

Corresponding Author: Mohammad Shirali-Shahreza, Department of Computer Science, Sharif University of Technology, Azadi Street, P.O. Box 11365-9415, Tehran, Iran Tel: 0098-912-2236349 Fax: 0098-21-66005117

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.

Most of the text steganography methods are designed for English texts and there are a few text steganography methods for other languages. In this study, we propose a new text Steganography method for Persian and Arabic texts. This method hides data in Persian and Arabic texts which are stored in Unicode format.

Our method is based on the fact that some letters in Persian and Arabic languages have different shapes in different places of words. This feature of Persian and Arabic languages is supported in Unicode format. This feature is used of Unicode standard for hiding data in Persian and Arabic texts.

A few works have been done on hiding information in texts. Following is the list of nine different methods of the works carried out and reported thus far for English text. After explaining these methods, the text Steganography methods that are especially designed for Persian and Arabic texts are surveyed.

Steganography of information in random character and word sequences (Bennett, 2004): By generating a random sequence of characters or words, specific information can be hidden in this sequence.

In this method, the characters or words sequence is random; therefore it is meaningless and attracts the attentions too much. It seems to be that this method is not steganography, but it is a kind of encryption.

Syntactic methods (Bennett, 2004): 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.

Line shifting (Low *et al*., 1995) and (Alattar and Alattar, 2004): 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.

Word shifting (Low *et al*., 1995) and (Kim *et al*., 2003): 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. Similar to Line Shifting method, retyping of the text or using OCR programs destroys the hidden information.

Semantic methods (Bennett, 2004): In this method, we use the synonym of words for 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 Line Shifting and Word Shifting methods). However, this method may alter the meaning of the text.

Feature coding (Rabah, 2004): In this method, some of the features of the text are altered. For example, the end part of some characters such as h, d, b or so on, are elongated or shortened a little thereby hiding information in the text. In this method, a large volume of information can be hidden in the text without making the reader aware of the existence of such information in the text.

By placing characters in a fixed shape, the information is lost. Retyping the text or using OCR program (as in Line Shifting and Word Shifting methods) destroys the hidden information.

Open spaces (Bender *et al*., 1996) and (Huang and Yan, 2001): 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.

Open spaces (Bender *et al.*, 1996) and (Huang and Yan, 2001): In this method, hiding information is done through adding extra white-spaces in the text. These white-spaces can be placed at the end of each line, at the end of each paragraph or between the words. This method can be implemented on any arbitrary text and does not raise attention of the reader.

However, the volume of information hidden under this method is very little. Also, some text editor programs automatically delete extra white-spaces and thus destroy the hidden information.

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.

Dot steganography method (Shirali-Shahreza and Shirali-Shahreza, 2006a): In the Dot steganography method (Shirali-Shahreza and Shirali-Shahreza, 2006a), 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 (Fig. 1), we hide information in the texts. This method does not attract attention and can hide a large volume of information in text.

La steganography method (Shirali-Shahreza, 2007): The La steganography method (Shirali-Shahreza, 2007) 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 be used on printed documents.

Pointed letters with extension method (Gutub and Fattani, 2007): The pointed letters with extension method (Gutub and Fattani, 2007), uses the pointed letters with extension (Kashida in Arabic) to hold secret bit one and

Fig. 1: Vertical displacement of the dots of the Persian letter NOON (Shirali-Shahreza, 2006a)

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

Fig. 2: Example of text steganography method proposed in by Gutub and Fattani (2007) (adding extensions after letters)

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. Figure 2 shows an example of this method.

Diacritic arabic method (Aabed *et al.*, 2007): In the diacritic Arabic method (Aabed *et al.*, 2007), 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 (Aabed *et al.*, 2007).

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.

MATERIALS AND METHODS

In this study, 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 (Shirali-Shahreza and Shirali-Shahreza, 2006b). Then we explain the Unicode Standard briefly and at last we explain our suggested method in full details.

The characteristics of Persian and Arabic: Arabic alphabet has 28 letters. Persian has all the letters of Arabic and four more letters of (Unicodes: 06AF, 0698, 0686, 0673). 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 0639 is written as FECB at the beginning of a word, as FECC in the middle, as FECA at the end and as FEC9 in the isolated form. We use this characteristic of Arabic and Persian languages in our method.

In Persian and Arabic the letters are connected to each other in writing, while in English the letters are written separately.

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 have dot (.) i and j.

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. The ZWJ's Unicode is U+200D. We use this character of Arabic and Persian languages in our method.

Unicode Standard: The Unicode Standard (The Unicode Consortium, 2006) 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 systems 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.

In the Unicode Standard, each Persian or Arabic letters has its unique code. Also, all shapes of each letter have their own code. For example, the code of letter Seen (س) in the Unicode Standard is 069B and the codes of different forms are FEB1 for the isolated form (س), FEB2 for the final form (س), FEB3 for the initial form (سـ) and FEB4 for the medial form (ـسـ).

For saving the documents in the Unicode Standard, only the unique code of each character is saved and the program which shows the letter will show the correct shape of letter regarding to its position in the word.

Our method: As described earlier, each Persian or Arabic letter can have four different shapes regarding to its position in the word and each Persian or Arabic letter have one unique code which show the letter in isolated form act as a word representative. But the four possible shape of letter including the isolated form (the initial form, the medial form, the final form and the isolated form) have their separate code in the Unicode Standard.

In the Unicode Standard, only the code of representative form of letters are saved in the text file and the program which shows the letter will show the correct shape of letters regarding to their position in the word.

However, one can save the text in Unicode Standard by inserting the code of correct shape of letters (regarding to their position in the word) instead of their representative letter code. Therefore, the text viewer -the program which shows the letter - does not determine the word shape automatically and only show the letter shape which is related to the saved code in the text.

The method proposed in this study for hiding data in Persian and Arabic Unicode texts is using this feature of the Unicode Standard.

For each letter in the text, we can save it by using the representative letter. But we can also save the letter by using the code of correct shape of the letter (regarding to its position in the word).

For hiding bit 0 in the word, the first option is used for saving the word. But for hiding bit 1 in the word, the second option is used.

But when we use the mixture of representative letters code and the code of shape of the letters in one word together, the text viewer does not select the correct shape of representative letters automatically and shows them in isolated form.

For solving this problem, we insert the ZWJ character between the two letters to connect them together. Because this character is a non-printing character, therefore, this method does not make any apparent changes in the original text and have a perfect perceptual transparency. A sample of the process of hiding data in a word is shown in Table 1.

For extracting data from stego text (the text contains hidden data), the code of letters is checked. If the letter is representative letter, we conclude that bit 0 is hidden, but if the letter code is its shape code (not the code of representative letter), we conclude that bit 1 is hidden in the letter. By putting all the bits of 0 and 1 next to each other we can extract the information hidden in the text.

Our method has very high hiding capacity, because we hide one bit in each letter. Now we estimate the amount of data we can hide in a Persian text. Assume that there are k words in the document and each word has α letters in average. After each word, there is a space or a punctuation mark such as comma. So, the size of the text is $2k(\alpha+1)$ bytes because there are $(\alpha+1)$ characters for each word and each character need two bytes in the Unicode.

Table 1: Hiding 101 in word شام

	Word	Unicode Representation
Input text	شام	0645,0627,0634
Data to hide	101	
Replaced characters	شام	FEE1,0627,FEB7
Adding required ZWJ	شام	FEE1,0627,200D,FEB7
Stego text	شام	FEE1,0627,200D,FEB7

In our method, we hide one bits in each letter of the word. So, we hide α bits in each word in average and a total of $k\alpha$ bit in the document. Therefore the hiding capacity of our method as bit/kilobyte is:

$$\text{Hiding Capacity} = 1024 \times \frac{k\alpha}{2k(\alpha+1)}$$

$$= 512 \times \frac{\alpha}{(\alpha+1)} \text{ bit/KB}$$

The average number of letters in each word (α) is 3.5 in Persian (Shirali-Shahreza, 1996). So, we have:

$$\text{Hiding capacity} = 512 \times \frac{\alpha}{(\alpha+1)}$$

$$= 512 \times \frac{3.5}{3.5+1} \cong 398 \text{ bit/KB}$$

This means that our method can hide about 400 bits of information in each kilobyte of a Persian text.

RESULTS AND DISCUSSION

In out method, the information is hidden in Persian and Arabic texts using the Unicode Standard.

We tested our method on some Persian text files. We selected the resources which are used in our earlier Persian and Arabic text Steganography methods (Shirali-Shahreza and Shirali-Shahreza, 2006a; Shirali-Shahreza, 2007) 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.

Table 2 shows that we can hide about 400 bits in each kilobyte of text.

As it is seen in the Table 2, our method capacity is very high, especially in comparison with La steganography method (Shirali-Shahreza, 2007). In this method we hide a bit of information in each Persian and Arabic letter, but in the Dot steganography method a bit of information is hid in each letter with dot. So, Present

Table 2: Comparing the capacity of our method with the Dot Steganography and La Steganography methods

Newspaper	Website address	Text size (bit/kilobyte)	Our method (kilobyte) text capacity (bit)	Our method capacity ratio (bit/kilobyte)	Dot steganography method (Shirali-Shahreza and Shirali-Shahreza, 2006a) capacity ratio	La steganography method (Shirali-Shahreza, 2007) 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

method capacity is four times higher than Dot steganography method (Shirali-Shahreza and Shirali-Shahreza, 2006a) in average.

Also, our method has advantages over these methods. For example, contrary to the Dot steganography method (Shirali-Shahreza and Shirali-Shahreza, 2006a), this method does not change the apparent of the text and does not required specific font.

CONCLUSIONS

In this study, a new method for Steganography in Persian and Arabic texts has been presented. This method uses the Unicode Standard and the special feature of Persian and Arabic languages that each letter has different shapes.

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.

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 (Cvejic, 2004).

Our method satisfies both perceptual transparency and hiding capacity requirements. It does 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/letter in the text file, so our method has high capacity

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 is not changed 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 the Arabic and Persian languages have other specific characteristics which 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 (Brassil *et al*., 1994).

REFERENCES

Aabed, M.A., S.M. Awaideh, A.M. Elshafei and A.A. Gutub, 2007. Arabic diacritics based steganography. Proceedings of the International Conference on Signal Processing and Communications (ICSPC 2007), November 24-27, Dubai, UAE, pp: 756-759.

Alattar, A.M. and O.M. Alattar, 2004. Watermarking electronic text documents containing justified paragraphs and irregular line spacing. Security, Steganography, and Watermarking of Multimedia Contents VI, Proceedings of SPIE, 5306, January 19, San Jose, CA, USA., pp: 685-695.

Bender, W., D. Gruhl, N. Morimoto and A. Lu, 1996. Techniques for data hiding. IBM Syst. J., 35: 313-336.

Bennett, K., 2004. Linguistic steganography: Survey, analysis and robustness concerns for hiding information in text, Purdue University. CERIAS Technical Report 2004-13. https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/2004-13.pdf.

Brassil, J.T., S. Low, N.F. Maxemchuk and L. O'Gorman, 1994. Marking text features of document images to deter illicit dissemination. Proceedings of the 12th IAPR International Conference on Pattern Recognition, October 9-13, IAPR, pp: 315-319.

Brassil, J.T., S. Low and N.F. Maxemchuk, 1999. Copyright protection for the electronic distribution of text documents. Proceedings of the IEEE, July 1999, IEEE Xplore London, pp: 1181-1196.

Cvejic, N., 2004. Algorithms for Audio Watermarking and Steganography. 1st Edn., Oulu University Press, Finland, ISBN: 951-42-7384-2 .

Gutub, A. and M. Fattani, 2007. 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), May 25-27, Vienna, Austria, pp: 28-31.

Hopper, N.J., 2004. Toward a theory of steganography. Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. http://reports-archive.adm.cs.cmu.edu/anon/anon/usr0/ftp/2004/CMU-CS-04-157.pdf.

Huang, D. and H. Yan, 2001. Interword distance changes represented by sine waves for watermarking text images. IEEE. T. Circ. Syst. Vid., 11: 1237-1245.

Kim, Y., K. Moon and I. Oh, 2003. A text watermarking algorithm based on word classification and inter-word space statistics. Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03), August 3-6, IAPR, pp: 775-779.

Low, S.H., N.F. Maxemchuk, J.T. Brassil and L. O'Gorman, 1995. Document marking and identification using both line and word shifting. Proceedings of the 14th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'95), April 2-6, IEEE, pp: 853-860.

Maxemchukand, N.F. and S. Low, 1997. Marking text documents. Proceedings of the IEEE International Conference on Image Processing, October 26-29, Santa Barbara, CA, USA., pp: 13-16.

Rabah, K., 2004. Steganography. The art of hiding data. inform. Technol. J., 3: 245-269.

Shirali-Shahreza, M.H., 1996. Off-line recognition of farsi handwritten words and numerals by neural networks. Ph.D. Dissertation, University of Technology (Tehran Polytechnic), Tehran, Iran.

Shirali-Shahreza, M.H. and M. Shirali-Shahreza, 2006a. A new approach to Persian/Arabic text steganography. Proceedings of the 5th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2006), July 10-12, Honolulu, HI, USA., pp: 310-315.

Shirali-Shahreza, M.H. and M. Shirali-Shahreza, 2006b. Persian/Arabic captcha. Iadis Int. J. Comput. Sci. Inf. Syst., 1: 63-75.

Shirali-Shahreza, M., 2007. 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), December 3-12 2007, Bridgeport, CT, USA. http://www.cisse2007.org/.

The Unicode Consortium, 2006. The Unicode Standard 5.0. 5th Edn., Addison-Wesley Professional, ISBN: 0321480910 .