A New Approach to Persian/Arabic Text Steganography

M. Hassan Shirali-Shahreza
Computer Eng. 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

Conveying information secretly and establishing hidden relationship has been of interest since long past. Text documents have been widely used since very long time ago. Therefore, we have witnessed different method of hiding information in texts (text steganography) since past to the present. In this paper we introduce a new approach for steganography in Persian and Arabic texts. Considering the existence of too many points in Persian and Arabic phrases, in this approach, by vertical displacement of the points, we hide information in the texts. This approach can be categorized under feature coding methods. This method can be used for Persian/Arabic Watermarking. Our method has been implemented by JAVA programming language.

Keywords: Information Security, Text Steganography, Text Watermarking, Information Hiding, Feature Coding, Persian/Arabic Text, Image Processing & Pattern Recognition.

1. Introduction

By development of computer and the expansion of its use in different areas of life and work, the issue of security of information has gained special significance. One of the concerns in the area of information security is the concept of hidden exchange of information. For this purpose, various methods including cryptography, steganography, coding and so on have been used. Steganography is one of the methods which have attracted more attention during the recent years.

In implementing steganography, the main objective is to hide the information under cover media so that the outsiders may not discover the information contained in the said frame. 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 [1].

Most of steganography works have been carried out on pictures [2, 3], video clips [4, 5], music and sounds [6].

Text steganography is the most difficult kind of steganography [7]; this is due largely to the relative lack of redundant information in a text file as compared with a picture or a sound file [8].

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.

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 [9].

Contrary to other media such as pictures, sounds and video clips, using text documents has been common since very old times. Even after invention of printing machine, most of the books and documents have contained only texts. This has extended until today and still, using text is preferred over other media, because the texts occupy lesser memory, communicate more information and need less cost for printing as well as some other advantages.

Proceedings of the 5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse (ICIS-COMSAR'06)
0-7695-2613-6/06 $20.00 © 2006 IEEE

As the use of text and hidden communication goes back to antiquity, we have witnessed to steganography of information in texts since past. For example, in order to prevent disclosure of government documents by the press, Margaret Thatcher, former British Prime Minister used to place certain number of white-spaces in documents related to each cabinet minister so that she could identity the owner of the document [10].

Today, the computer systems have facilitated hiding information in texts. The range of using hiding information in text has also developed. From among the most important of these technologies, one can name of hiding information in electronic texts and documents. The use of hiding information in text for web pages is another example.

Different methods are used for hiding information in text which will be dealt with in section 2.

The present paper offers a new method for hiding information in Persian and Arabic texts. Due to differences between languages, no single method can be used for hiding information in texts of different languages. This will be discussed in section 3.

2. Previous Works

A few works have been done on hiding information in texts. Following is the list of ten different methods of the works carried out and reported thus far.

2.1. Steganography of Information in Random Character and Word Sequences [11]

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.

2.2. Steganography of Information in Specific Characters in Words [10]

In this method, some specific characters from certain words are selected as hiding place for information. In the simplest form, for example, the first words of each paragraph are selected in a manner that by placing the first characters of these words side by side, the hidden information is extracted. This has been done by classic poets of Iran as well.

This method requires strong mental power and takes a lot of time. It also requires special text and not all types of texts can be used in this method.

2.3. Creating Spam Texts [11]

Another feature of HTML documents is their case-insensitivity of tags and their members. For example, the three tags
,
 and
 are equally valid and are the same. As a result, one can do information steganography in HTML documents by changing the small or large case of letters in document tags. To extract information, one can extract information by comparing these words with words in normal case and by using the appropriate function.

However, in the WML, all tags should be written in lowercase letters and, as a result, this method cannot be employed.

2.4. Line Shifting [12, 13]

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.

2.5. Word Shifting [12, 14]

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 method described under 2-4, retyping of the text or using OCR programs destroys the hidden information.

2.6. Syntactic Methods [11]

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

Proceedings of the 5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse (ICIS-COMSAR'06)
0-7695-2613-6/06 $20.00 © 2006 IEEE

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

2.7. Semantic Methods [11, 15]

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 methods listed under 2-4 and 2-5).

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

2.8. Feature Coding [16]

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 methods 2-4 and 2-5) destroys the hidden information.

2.9. Abbreviation [8]

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.

2.10. Open Spaces [8, 17]

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.

3. Suggested Algorithm

One of the characteristics of Persian language is abundance of points in its letter. Although English also has points, there is a huge difference between the two languages in this respect. In English, only two letters of small "i" and small "j" have point while in Persian 18 letters out of 32 alphabet letters have points. From these 18, 3 letters have 2 points each, 5 letters have 3 points each and the other 10 letters have 1 point each (Table 1) [18].

Persian language, of course, has 4 letters make than Arabic language from which, 3 letters have point. Therefore, in Arabic, 15 letters out of the entire 28 alphabet letters have point. In general, the number of points in any given Persian or Arabic text is noteworthy. In this paper, this same characteristic of Persian and Arabic languages is used for steganography.

For this purpose, the concerned information is first of all compressed. Then, we look for the first pointed letter in the given text. By finding this character, we go to the compressed information and read the first bit of information which has one the values of zero or one. If the value of the bit were zero, the concerned character remains unchanged. If the value of the bit were one, we shift the point on the concerned character a little upward (Figure 1).

Table 1. Persian Alphabet

ا ح د ر س ص ط ع ک گ ل م و ه	ب ج خ ذ ز ض ظ غ ف ن	ت ق ی(ي)	پ ث چ ژ ش
Letters without point	Letters with one point	Letters with two points	Letters with three points

Figure 1. Vertical displacement of the points for the Persian letter NOON

This procedure is repeated for the next pointed characters in the text and the next bits of information. Thus, the entire information is hidden. In order to divert the attention of readers, after hiding all information, the points of the remaining characters are also changed randomly. Of course, before doing this, the size of hidden information is also hidden in the beginning of the text.

For the characters with two or three points, all points shift, because shifting one point among the points of a character raises attentions.

While extracting information from the text, the program starts identifying the quantity of hidden bit in the character based on the place of points on the character. By placing all the extracted bits side by side, the compressed information is obtained. Now, this compressed piece of information is uncompressed and the original data is recovered.

This method can be categorized under feature coding methods (described under 2.8) which is developed for Persian and Arabic languages.

4. Advantages and Disadvantages

4.1. Advantages

1. By this method, a large volume of information can be hidden in text, because a large number of letters in Persian and Arabic have points.
2. Due to the lack of a strong OCR program for Persian and Arabic languages, the printed text cannot be easily converted into a simple text thus destroying the hidden information is difficult.
3. The text containing hidden phrases is not specific to computer and the hidden information can also extracted from printed text. In order to recover the information in case of printed text, the text should be scanned and then subjected to the relevant program.
4. The hidden text is resistant to enlargement or downsize and these changes do not destroy the hidden information.

4.2. Disadvantages

1. The information is lost in case of retyping.
2. The output text has a fixed frame due to the use of only one font.
3. Due to the lack of good OCR program for Persian and Arabic languages, using this method in texts that are printed and then scanned is difficult.

5. Experimental Results

In this project, files and information were hidden in text documents by the use of the described algorithm. For this purpose, several files containing text, picture and executive file were selected. Then the files were compressed to reduce their size. The compressed file was then hidden in the text by our steganography program.

The steganography program is developed by Java language. In order to hide information in the text, first of all we introduce some change to the concerned font. That is, we define a new mode for the pointed letters in blank spaces of the font file. In this mode, the point of the letter is placed a little higher. Now the program starts to read the incoming data bit by bit and the incoming text letter by letter to find pointed letters. If the incoming bit was zero, the pointed letter remains unchanged but if the incoming bit were one, the found pointed letter is changed to a mode whose point is a little higher. Then the resulted output file is converted into a PDF file by Adobe Acrobat. The concerned document is now ready and can be printed.

In order to extract data from the text, we used information extractor program which is also in Java language. For this purpose, first we converted the PDF file to a framed text file by the use of Acrobat Reader. Then, by the use of the algorithm for extraction of information, the program acted reversely and extracted the compressed data from the text. At the end, we uncompressed the data file to obtain the original file.

By comparing the output files with input files, we observed that both files were identical.

After this experiment, we studied the capacity of a number of texts selected from some highly circulating newspapers of Iran for hiding data. In steganography the hidden data must not attract attention. Therefore for using newspapers or magazines as cover media for hiding data in them, it is better to use internal pages instead of using first page or cover pages. In this project, we check sport pages of some Iranian newspapers for computing the capacity of an article for hidden data. Table I shows the result of this computation.

The internet address of these newspapers and the capacity of each text for hiding data are shown in Table 2. All the articles selected on 20 August 2005.

Table 2. The capacity of an article in sport pages of some Iranian newspapers for hiding data

Newspaper	WebSite Address	Text Size (Kilo Byte)	Text Capacity(Bit)	Capacity Ratio (Bit / Kilobyte)
Farhange Ashti	ashtidaily.com	13.3	1278	96
Hamshahri	hamshahri.net	6.82	820	120
Iran	iraninstitute.org	6.64	694	105
JameJam	jamejamdaily.net	3.84	434	113
Javan	javandaily.com	8.03	922	115
Jomhouri Eslami	jomhourieslami.com	3.52	441	125
Keyhan	kayhannews.ir	2.92	310	106
Khorasan	khorasannews.com	5.40	628	116
Quds	qudsdaily.net	9.98	1137	114
Shargh	sharghnewspaper.com	20.4	2409	118

6. Conclusion

In this paper we introduce a new approach for steganography of information in Persian and Arabic texts. This method is based on the existence of point in majority of letters of Persian and Arabic alphabets. On this basis, information was hidden in text by changing the place of points. This method can be used in hidden exchange of information through text documents and text watermarking.

In addition to establishing secret communication, this method can be used for preventing illegal duplication and distribution of texts especially electronic texts [19, 20].

In addition to use this method for electronic texts, it can be applied on hard copy documents. To this end, we print the document after hiding data in it. For extracting data from the hard copy document, we scan it and unhide the embedded data by computer.

Considering the similarity of Urdu script (official language of Pakistan) with Persian and Arabic, this method can be used in Urdu as well.

In addition to vertical displacement of points, the points can be displaced in horizontal direction as well and thus two bits of information can be hidden in each letter. By combining the above method with other methods such as line shifting and word shifting, the volume of hidden information can be increased.

By employing a font editing software, the program can be enabled dynamically to produce necessary fonts for hiding information so that the output form of the text is not homogenous and conform to the input form of the text.

7. References

[1] J.C. Judge, "Steganography: Past, Present, Future", *SANS white paper*, November 30, 2001, http://www.sans.org/rr/papers/index.php?id=552, last visited: 1 May 2006.

[2] R. Chandramouli, and N. Memon, "Analysis of LSB based image steganography techniques", *Proceedings of the International Conference on Image Processing*, vol. 3, 7-10 Oct. 2001, pp. 1019 - 1022.

[3] M. Shirali Shahreza, "An Improved Method for Steganography on Mobile Phone", *WSEAS Transactions on Systems*, vol. 4, Issue 7, July 2005, pp. 955-957.

[4] G. Doërr and J.L. Dugelay, "A Guide Tour of Video Watermarking", *Signal Processing: Image Communication*, vol. 18, Issue 4, 2003, pp. 263-282.

[5] G. Doërr and J.L. Dugelay, "Security Pitfalls of Frame-by-Frame Approaches to Video Watermarking", *IEEE Transactions on Signal Processing*, Supplement on Secure Media, vol. 52, Issue 10, 2004, pp. 2955-2964.

[6] K. Gopalan, "Audio steganography using bit modification", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '03)*, vol. 2, 6-10 April 2003, pp. 421-424.

[7] J.T. Brassil, S. Low, N.F. Maxemchuk, and L. O'Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying", *IEEE Journal on Selected Areas in Communications*, vol. 13, Issue. 8, October 1995, pp. 1495-1504.

[8] 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.

[9] N. F. Maxemchuk and S. Low, "Marking Text Documents", *Proceedings of the IEEE International Conference on Image Processing*, Santa Barbara, CA, USA, Oct. 26-29, 1997, pp. 13-16.

[10] T. Moerland, "Steganography and Steganalysis", May 15, 2003, www.liacs.nl/home/tmoerlan/privtech.pdf, last visited: 1 May 2006.

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

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

0-7695-2613-6/06 $20.00 © 2006 IEEE

[13] 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.

[14] 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

[15] 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, July 3-4, 2003.

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

[17] 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

[18] M. H. Shirali-Shahreza, and S. Shirali-Shahreza, "A Robust Page Segmentation Method for Persian/Arabic Document", *WSEAS Transactions on Computers*, vol. 4, Issue 11, Nov. 2005, pp. 1692-1698.

[19] J.T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright Protection for the Electronic Distribution of Text Documents", *Proceedings of the IEEE*, vol. 87, Issue. 7, July 1999, pp. 1181-1196.

[20] J. T. Brassil, S . Low, N. F. Maxemchuk, and L. O'Gorman, "Marking Text Features of Document Images to Deter Illicit Dissemination", *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, 1994, vol. 2, 9-13 Oct. 1994, pp. 315-319.