

A New Text Steganography Method By Using Non-Printing Unicode Characters

Akbas E. Ali *

Received on: 5/5/2009

Accepted on: 1/10/2009

Abstract

One of the methods used in security areas is steganography. Steganography is the art and science of hiding information by embedding messages within cover media without attracting attention. the cover media can be text, image, video or audio files. Text steganography is more difficult than others due to the difficulty in finding redundant information in text file. This paper presents a new idea for text steganography by using Unicode standard characters, (which have the non-printing properties) to encode the letters of English language and embedding the secret message letter by letter into the cover-text.

This method has high hiding capacity, it can hide (K+1) letters in a text with K characters and it does not make any apparent changes in the original text. So it satisfies perceptual transparency.

Keywords: text steganography, Unicode standard, non-printing characters, perceptual transparency

طريقة جديدة للاخفاء بالنص باستخدام حروف يونيكود غير مرئية

الخلاصة

الاختزال (الاخفاء) هو احد الطرق المستخدمة لحماية المعلومات. فن الاختزال هو علم اخفاء المعلومات وذلك بتضمين رسائل في وسط معلوماتي يستخدم كغطاء بدون جذب الانتباه. الغطاء ممكن ان يكون ملف يحتوي على نص او صورة او فيديو. الاختزال في النص هو الاصعب وذلك لصعوبة ايجاد معلومات فائضة في ملف النص. هذا البحث يقدم فكرة جديدة للاختزال في النص باستخدام حروف من نظام يونيكود (والتي تتصف بأنها غير مرئية عند الطباعة) استخدمها لترميز حروف اللغة الانكليزية ثم اخفاء الرسالة السرية حرف حرف في ملف الغطاء. هذه الطريقة تمتلك سعة اخفاء عالية حيث تستطيع اخفاء (K+1) من الحروف في نص يحتوي على K من الحروف وكذلك فانها لا تسبب اي تغيير في شكل النص الاصيلي اي انها تحقق شفافية عالية

1. Introduction

Privacy is a major concern for users of public networks such as the Internet. Traditionally, privacy is among the central concerns of cryptography, which achieves private communication

through encryption. One problem with encryption, however, is that although it may hide the contents of a message, the mere transmission of an encrypted message may reveal something about that message's contents [7].

Steganography is the art and science of hiding information by embedding messages within other, seemingly harmless messages called the carrier. The word steganography literally means covered writing as derived from Greek. It includes a vast array of methods of secret communications that conceal the very existence of the message. Among these methods are invisible inks, microdots, Character arrangement (other than the cryptographic methods of permutation and substitution), digital signatures, covert channels and spread-spectrum communications. Most of steganography works have been carried out on pictures, video clips, music and sounds. Text steganography is the most difficult kind of steganography; this is due largely to the relative lack of redundant information in a text file as compared with a picture or a sound file [6].

The steganography process generally involves placing a hidden message in some transport medium, called the carrier. The secret message is embedded in the carrier to form the steganography medium. The use of a steganography key may be employed for encryption of the hidden message and/or for randomization in the steganography scheme. In summary [1]:

Cover_medium+hidden_data+ stego_key = stego_medium

In this context, the cover_medium is the file in which we will hide the hidden_data, which may also be encrypted using the stego_key. The resultant file is the stego_medium (which will, of course, be the same type of file as the cover_medium). The cover_medium (and, thus, the stego_medium) are typically text, image or audio files.

2. Previous works:

A few works have been done on hiding information in texts. One of these works is the paper titled with “ High capacity Persian/Arabic text steganography, (Shirali-Shahreza, 2008) “.In this paper a new method is proposed for hiding information in Persian and Arabic Unicode texts.

In Arabic letters based languages, some letters are connected together in a word (this feature is not exist in roman letters based languages). To solve the problem of this feature we use two characters in the Unicode standard. They are zero width non joiner (ZWNJ) and zero width joiner (ZWJ) characters, which are prevents the Arabic letters from joining or forces them to join together respectively.

In this method by using these two special characters, the information is hidden in Persian and Arabic Unicode text documents. The hiding capacity

of this method is one bit in each letter.

Our proposed method is to use these two special characters (ZWNJ and ZWJ) for hiding the information in roman letters based languages and to increase the hiding capacity more than the previous mentioned method by embedding one character a time instead of one bit.

3. Materials and methods

In this paper, we used the Unicode standard to represent text to be hidden. We need to introduce a brief explanation about the Unicode standard and its manipulation with some features of Arabic writing system.

3.1 Unicode standard:

Unicode is rapidly becoming the preferred means for representing symbols used in creating multimedia content, especially for information that's presented in multiple languages [4]. It is the standard for digital representation of the characters used in writing all of the world's languages. Unicode provides a uniform means for storing, searching, and interchanging text in any language. It is used by all modern computers and is the foundation for processing text on the Internet. Unicode is developed and maintained by the Unicode Consortium: <http://www.unicode.org/>[8].

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, Cyrillic (Russian and Serbian), Greek, Arabic (including Arabic, Kurdish, Persian and Urdu), Hebrew, India, Armenian, Assyrian, Chinese, Katakana, Hiragana (Japanese) and Hangeul (Korean). Moreover there are a lot of mathematical and technical symbols, punctuation marks, arrows and miscellaneous.

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 local guide [5].

3.2. Unicode standard and feature of Arabic language:

In the Unicode standard, each 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 (س) and FEB4 for the medial form (س). 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 [5]

Instead of using the four possible shapes of Arabic letters (including the initial form, the medial form, the final form and the isolated form) the Unicode standard provide a unique code which shows the letter in isolated form act as a word representative. This representative letter can be used with another non-printing code which will give the required shape of the letters, so for each letter in the text, we can save it by using the representative form of letter mixed with the code of correct shape of the letter (regarding to its position in the word) [5]. These non-printing

characters used to connects Arabic language letters in the required shape, they are:

1. The Zero width Joiner (Zwj) : used to connect two character, (Unicode = U+200D)
2. The Zero width non-Joiner (ZWNJ): used to disconnect two characters,(Unicode = U+200C).

3.3 Proposed hiding method:

In this paper, a new steganography scheme is introduced for hiding text in a cover-text. The proposed method based on the non-printing Unicode characters ZWJ and ZWNJ, which are used for joining and disjoining the Arabic letters, will be used for another purpose in the roman letters based texts. Here, the technique will be used just for hiding the information, where the embedding of ZWJ in the cover-text will represent "1" and embedding of ZWNJ will represent "0".

Further, the embedding capacity will be increased by inserting a group of this non-printing Unicode characters to represent one letter of English language. By this approach the embedding process will insert one letter at each time instead of hiding one bit a time.

Codes with four digits at most will be enough to represent the 26 English letters; these codes are illustrated in table-1:

In table-1; each English letter had been allotted with a binary code, and then, this binary code is converted into a Unicode representation, depending on the

assumption that each “1” and “0” of the binary code is converted into ZWJ (U+200D) and ZWNJ (U+200C) respectively. As an example, the binary code of the letter “A” is “0” and the corresponding Unicode representation is “200C”. The same thing for the letter “Z” which takes the binary code of “1011”, and the corresponding Unicode representation is “200D,200C,200D,200D”.

The Unicode collection representation, which are illustrated in table-1 could be inserted, to represent the letters of the secret message between any two characters of the cover-text according to algorithm-1, and the receiver of the stego-text can extract the secret message by algorithm-2.

A sample of the process of embedding data in text is shown in table-2 and table-3

4. Result and discussion

- In our method, the information can be hidden in English Unicode texts using both ZWNJ and ZWJ characters. This approach will insert one letter at each time instead of hiding one bit. So the hiding capacity will be increased eight times.

- The proposed method was tested on some text files to compute the capacity of the hiding. It had been seen that one letter can be hidden before and after any character of the cover-text, so (K+1) letters can be hidden in text with K characters.

- The using of the non-printing

Unicode characters (ZWJ and ZWNJ) does not make any apparent changes in the plain original text. Except the increasing of the size of the output stego-text file, the original text does not change by the hiding process therefore this method has a perfect perceptual transparency.

5. Future Works

Text which is written in different languages or other types of information: image, audio can be hidden in text by the same approach.

To improve the quality of the security of data, the secret message can be encrypted before hidden.

The use of a steganography-key may be employed for randomization of the embedding process.

6. Conclusion

A new text steganography method is presented in this paper. This method uses the non-printing Unicode characters (ZWJ and ZWNJ) as a tool to represent codes for English letters.

The stego Unicode texts will not change during copy and paste between computer programs. And this method is not dependent on any special format of text, so it is suitable to use in HTML web pages, Microsoft PowerPoint and Word, etc.

The proposed method satisfies both hiding capacity requirement and perceptual transparency; it can hide (K+1) English language

letters in cover-text with K letters and it does not make any apparent changes in the original text. The method can hide any type of information (audio, image) in text.

7. References

- [1] Kessler, c. Gary, “An Overview of Steganography”, the Computer Forensics Examiner issue of Forensic Science Communications, July 2004
- [2] Julie Kremer, “Steganography”, <http://www.nku.edu/~mcsc/mat494/uploads/Steganography.pdf>
- [3] Salomon, D., “Data privacy and security Encryption and information Hiding”, Springer, 2003
- [4] FJ Mabry and JR James and AJ Ferguson, “Unicode Steganographic Exploits: Maintaining Enterprise Border Security”, IEEE Security & Privacy, Vol. 5, No. 5., September 2007, pp. 32-39.
- [5] M.Shirali-Shahreza and S.Shirali-Shahreza, “high capacity Persian/Arabic text steganography”, Journal of applied sciences 8 (22): 4173-4179, 2008
- [6] Memon, jibrana. and khowaja, k. and K. Hameedullah, “EVALUATION OF STEGANOGRAPHY FOR URDU /ARABIC”, Journal of Theoretical and Applied Information Technology, 2008
- [7] Samphai boon, N. and Matthew N. Dailey, “Steganography in Thai Text”, Thailand, 2007.
- [8] Unicode Consortium: home page <http://www.unicode.org/>

Table (1) Codes for English letters

letter	Binary Code	Unicode collection representation	Letter	Binary Code	Unicode collection representation
A	0	200C	N	101	200D,200C,200D
B	1	200D	O	0000	200C,200C,200C,200C
C	00	200C,200C	P	0001	200C,200C,200C,200D
D	10	200D,200C	Q	0010	200C,200C,200D,200C
E	01	200C,200D	R	0100	200C,200D,200C,200C
F	11	200D,200D	S	1000	200D,200C,200C,200C
G	000	200C,200C,200C	T	0011	200C,200C,200D,200D
H	001	200C,200C,200D	V	0110	200C,200D,200D,200C
I	011	200C,200D,200D	U	1100	200D,200D,200C,200C
J	111	200D,200D,200D	W	1001	200D,200C,200C,200D
K	010	200C,200D,200C	X	0111	200C,200D,200D,200D
L	100	200D,200C, 200C	Y	1110	200D,200D,200D,200C
M	110	200D,200D,200C	Z	1011	200D,200C,200D,200D

Table (2) embedding the word “computer” in the cover-text “science”

	text	Unicode collection representation
Cover-text	science	0073,0063, 0069, 0065,006E,0063,0065
Data to hide	C	200C 200C
	O	200C 200C 200C 200C
	M	200D 200D 200C
	P	200C 200C 200C 200D
	U	200D 200D 200C 200C
	T	200C 200C 200D 200D
	E	200C 200D
	R	200C 200D 200C 200C
Stego-text	Science	200C, 200C, <u>0073</u> , 200C, 200C, 200C, 200C, <u>0063</u> , 200D, 200D, 200C, <u>0069</u> , 200C, 200C, 200C, 200D, <u>0065</u> , 200D, 200D, 200C, 200C, <u>006E</u> , 200C, 200C, 200D, 200D, <u>0063</u> , 200C, 200D, <u>0065</u> , 200C, 200D, 200C, 200C

Table (3) Second Embedding Example

Cover- text file: Size=26,112 bytes		The University Of Technology was established in 1975							
The Unicode collection representation of the cover-text will be:									
T	0054	i	0069	n	006E	e	0065		0020
h	0068	tI	0074	o	006F	s	0073	i	0069
e	0049	y	0079	l	006C	t	0074	n	006E
	0020		0020	o	006F	a	0061		0020
U	0055	o	006F	g	0067	b	0062	1	0031
n	006E	f	0066	y	0079	l	006C	9	0039
i	0069		0020		0020	i	0069	7	0037
v	0076	T	0054	w	0077	s	0073	5	0035
e	0065	e	0065	a	0061	h	0068		
r	0072	c	0063	s	0073	e	0065		
sI	0073	h	0068		0020	d	0064		
secret message		The main objective in steganography is to hide the information							
The Unicode collection representation of the secret message will be:									
T	200C,200C,200D,200D	s	200D,200C,200C,200D	i	200C,200D,200D				
h	200C,200C,200D	t	200C,200C,200D,200D	d	200D,200C				
e	200C,200D	e	200C,200D	e	200C,200D				
m	200D,200D,200C	g	200C,200C,200C	t	200C,200C,200D,200D				
a	200D	a	200C	h	200C,200C,200D				
i	200C,200D,200D	n	200D,200C,200D	e	200C,200D				
n	200D,200C,200D	o	200C,200C,200C,200C	i	200C,200D,200D				
o	200C,200C,200C,200C	g	200C,200C,200C	n	200D,200C,200D				
b	200C	r	200C,200D,200C,200C	f	200D,200D				

j	200D,200D,200D	a	200C	o	200C,200C,200C,200D
e	200C,200D	p	200C,200C,200C,200D	r	200C,200D,200C,200D
c	200C,200C	h	200C,200C,200D	m	200D,200D,200C
t	200C,200C,200D,200D	y	200D,200D,200D,200C	a	200C
i	200C,200D,200D	i	200C,200D,200D	t	200C,200C,200D,200D
v	200D,200C,200D	s	200D,200C,200C,200C	i	200C,200D,200D
e	200C,200D	t	200C,200C,200D,200D	o	200C,200C,200C,200D
i	200C,200D,200D	o	200C,200C,200C,200C	n	200D,200C,200D
n	200D,200C,200D	h	200C,200C,200D		

stego- text file: Size=26,624 bytes	The University Of Technology was established in 1975.				
The Unicode representation of the stego-text will be:					
200C,200C,200D,200D	0054	200C,200C,200D	0068	200C,200D	0049
200D,200D,200C	0020	200C	0055	200C,200D,200D	006E
200D,200C,200D	0069	200C,200C,200C,200D	0076	200D	0065
200D,200D,200D	0072	200C,200D	0073	200C,200C	0069
200C,200C,200D,200D	0074	200C,200D,200D	0079	200D,200C,200D	0020
200C,200D	006F	200C,200D,200D	0066	200D,200C,200D	0020
200D,200C,200C,200D	0054	200C,200C,200D,200D	0065	200C,200D	0063
200C,200C,200C	0068	200C	006E	200D,200C,200D	006F
200C,200C,200C,200D	006C	200C,200C,200C	006F	200C,200D,200C,200D	0067
200C	0079	200C,200C,200C,200D	0020	200C,200C,200D	0077
200D,200D,200D,200D	0061	200C,200D,200D	0073	200D,200C,200C,200D	0020
200C,200C,200D,200D	0065	200C,200C,200C,200D	0073	200C,200C,200D	0074
200C,200D,200D	0061	200D,200C	0062	200C,200D	006C
200C,200C,200D,200D	0069	200C,200C,200D	0073	200C,200D	0068
200C,200D,200D	0065	200D,200C,200D	0064	200D,200D	0020
200C,200C,200C,200D	0069	200C,200D,200C,200D	006E	200D,200D,200C	0020
200C	0031	200C,200C,200D,200D	0039	200C,200D,200D	0037
200C,200C,200C,200D	0035	200D,200C,200D			

Table (4) The Unicode representation

	Unicode representation		Unicode representation		Unicode representation
a	0061	w	0077	S	0053
b	0062	x	0078	T	0054
c	0063	y	0079	U	0055
d	0064	z	007A	V	0056
e	0065	A	0041	W	0057
f	0066	B	0042	X	0058
g	0067	C	0043	Y	0059
h	0068	D	0044	Z	005A
i	0069	E	0045	0	0030
j	006A	F	0046	1	0031
k	006B	G	0047	2	0032
l	006C	H	0048	3	0033
m	006D	I	0049	4	0034
n	006E	J	004A	5	0035
o	006F	K	004B	6	0036
p	0070	L	004C	7	0037
q	0071	M	004D	8	0038
r	0072	N	004E	9	0039
s	0073	O	004F	(ZWNJ)	200C
t	0074	P	0050	(ZWJ)	200D
u	0075	Q	0051		
v	0076	R	0052		

Algorithm 1: embedding process

Input: cover-text & secret message

Output: stego-text to transmit

Repeat (for each letter in the secret message)

Convert next letter of the secret message into corresponding binary code, and then into its Unicode collection representation (from table 1)

Print this Unicode collection representation in stego-text file

Get next character from the cover-text and print it in stego-text file.

Until secret message is finished

End.

Algorithm 2: extracting process
Input: received stego-text
Output: secret message & cover-text
Repeat (for each letter in the cover-text)
Get Unicode collection representation from next character of the cover-text
Convert the resultant Unicode collection into its corresponding letter (table1)
Print the produced letter in secret message file
Until cover-text is finished
End.

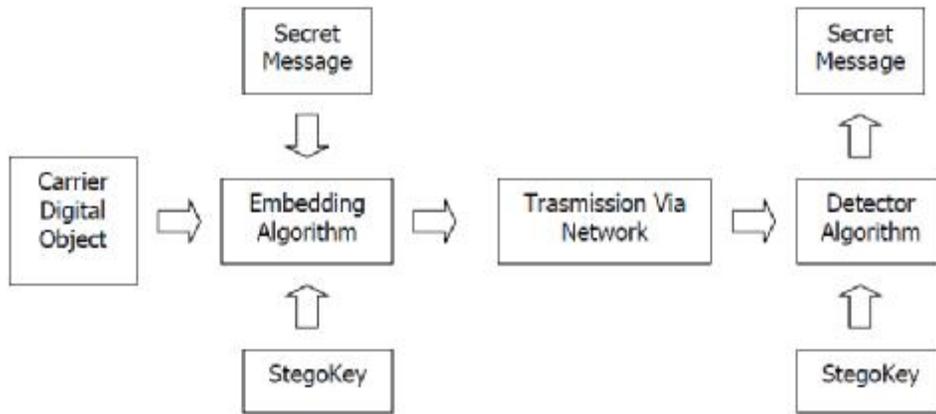


Figure (1) Model of data hiding technique [1]