An Email based high capacity text steganography scheme using combinatorial compression

Rajeev Kumar[1], Satish Chand[2], and Samayveer Singh[3]
Division of Computer Engineering,
Netaji Subhas Institute of Technology, New Delhi, India
rajeevgargnsit@gmail.com[1], schand20@gmail.com[2], and samayveersingh@gmail.com[3]

***Abstract*—In this paper, we propose an email based high capacity text steganography method using combinatorial compression. The method makes use of forward email platform to hide the secret data in email addresses. We use the combination of BWT + MTF + LZW coding algorithm to increase the hiding capacity, as it is proved that this combination increases the compression ratio. To further increase the capacity, the numbers of characters of email id are also used to refer the secret data bits. Furthermore, the method adds some random characters just before the '@' symbol of email ids to increase the randomness. Experimental results show that our method performs better than the some important existing methods in terms of hiding capacity.**

Keywords— text steganography; BWT; MTF; LZW;

I. INTRODUCTION

Steganography is a discipline of concealing secret messages with reliability in such a way that no one can be aware of the existence of the hidden messages. Another way to send the secret data securely is cryptography which encrypts or encodes the secret data rather than making it invisible. Thus, the advantage of steganography over cryptography alone is that the hidden messages do not attract any attention. Therefore, steganography can be said to protect the content of messages as well as the communicating parties. Information to be concealed and communicated covertly is called the payload or secret message. The carrier with concealed secret data is known as stego media or covert message. The cover media can be image, text, video or audio. The steganography method which uses text as cover media is known as the text steganography method. Text steganography is one of the toughest areas of data encryption, since the difference between the original and the covered texts is easily detectable. A good steganographic method must possess three properties: hiding capacity, security and robustness. Hiding capacity is the amount of secret data that can be concealed in the media. Security is related to the ability of a masquerader to figure the hidden information easily. Robustness is related to the resistance of the possible alteration to the unseen data [1].

In this paper, we propose an email based data hiding method which uses combinatorial compression to increase the hiding capacity. Basically, data compression algorithms are classified into two categories: lossless or lossy. Lossless data compression involves recovery of original data after decompression and lossy data compression losses some information while compression, hence exact recovery is not possible after decompression in case of lossy compression. In our proposed method, we use lossless data compression techniques as we are dealing with textual information. If the text contents are made even wee disturbed then also the meaning of the entire sentence can be changed. Here, we make use of Burrows Wheeler Transform (BWT) + Move to Front (MTF) encoding + LZW algorithms to achieve better compression ratio. The secret data is embedded into the email ids of forward mail platform. The cover text is chosen from the text base after some processing. While arranging the stego cover as a forward mail platform, the previously arranged email address list is utilized for selecting the email addresses. This email address list is used as a global stego key that is shared between both the sender and the receiver beforehand. To check the quality of our proposed method, we evaluate it on capacity metric. The rest of the paper is presented as follows. Section 2 summarizes the related works. Section 3 introduces the proposed work and in section 4, experimental results are analyzed. At last, in section 5, conclusion is provided.

II. RELATED WORKS

In this section, some of the important text steganography schemes for variety of languages like English, Persian Chinese, and Arabic, etc. are discussed.

Earlier in text steganography, Wayner [3,4] discussed an important method using mimic functions. It applies the inverse of Huffman Code having employed the randomly distributed bits of input stream on itself. It is directed toward the security perspective i.e. resiliency against statistical attacks. Another text steganography given by Maher [5] which is popular as TEXTO is built to transform uuencoded or PGP ASCII-armored ASCII data into English language sentences. It converts the secret data into English words. To extend the work of [6], another important method is synonyms-based approach [7-9]. Unlike [6], the method uses legitimate words and sentences having appropriate preciseness. Thus, the visual attack will have wee significance on these types of methods.. Sun et al. [10] introduced a method using the left and right components of Chinese characters hence known as L-R scheme. It selects characters with left and right components as candidates to conceal the secret data. If the secret data bit is "1", the scheme modifies the candidate by adjusting the space between the left and right components otherwise leaves unchanged. To improve the L-R scheme in terms of hiding capacity Wang et al. [11] extends this method by

978-1-4799-4236-7/14/$31.00 ©2014 IEEE

incorporating the up and down structure of Chinese characters as an extra candidate set. Further, a reversible function is also incorporated to get the original cover text after the initial hidden secret data has been extracted. Wang and Chang [12] discussed a text steganography method which hides the secret information into emotional icons (also known as emoticons) in chat rooms over the Internet. Collaboratively, both the sender and the receiver built a table. It is used at the time of communication. These emoticons are categorized into multiple classes according to their meaning (like smile laugh, cry). Therefore, each emoticon is fall into at-least one classes. The secret bits are used to be referred by the order number of an emoticon. Grothoff et al. [13] discusses a text steganography method which uses the errors to hide the secret data. The error is usually come in the way in a machine translation (MT). The secret data is hidden using substituting procedure on the translated text using translation variations of multiple MT systems. In 2008, Por et al. [14] discussed a text steganography method called WhiteSteg which hides secret data into the inter word and inter paragraph spacing. Normal space and tab space characters are mainly utilized to hide the secret data bits. the method hides significant amount of data into the spacing but suffers from the security point of view because the unusual ordering of the normal and tab spaces is easily seen by the show/hide tool. The normal space is indicated by dot sign and tab space by arrow sign. Por et al. [15] introduced a data hiding method based on space character manipulation called UniSpaCh. UniSpaCh conceals secret data in Microsoft Word document using Unicode space characters. It provides security from the show/hide attack. The method uses the eight Unicode characters which are not visible by show/hide option of word. The permutation of these Unicode characters is utilized to hide the secret data. Therefore, without making any significant alteration in the cover text spaces, UniSpaCh provides sufficient hiding capacity and to provide security some of the existing cryptography schemes are applied to secret data before embedding. Thus security is also improved.

Another important and popular method is introduced by Satir and Isik [16] using forward mail platform to hide the secret data. The method majorly considers hiding capacity and security issues of data hiding techniques in account. The LZW compression scheme is utilized to increase the hiding capacity. It tries to enhance the dual pattern repetition before employing the LZW scheme on the secret data. It hides the secret data into the email addresses which are listed in Carbon copy (Cc) field. In our proposed method, we extend the work of [16] to again increase the hiding capacity and security. Our method makes use of combination of Burrows Wheeler Transform (BWT) + Move to Front (MTF) encoding + LZW algorithms to increase the compression ratio. As [17] discussed that the Burrows Wheeler Transform (BWT), Move to Front (MTF) encoding, and LZW algorithms scheme used in combination as in this paper increases the compression ratio. In addition, we use the number of characters in email ids to refer to the secret data bits so that the hiding capacity is further increased. The next section discusses our proposed method.

III. THE PROPOSED METHOD

The goal of the proposed scheme is to hide more secret data and improve security so that the communication cost is reduced. The proposed scheme hides the secret data in text media. The method is divided into two phases: embedding and Extraction phase explained as follows.

3.1. Embedding phase

Let,

S: secret message

T: A text base of cover texts in which the secret message will be embedded.

K_1: set of email addresses having four characters before '@' symbol. It is shared between the sender and the receiver.

A: set of the second parts of email addresses such as outlook.com, gmail.com, etc.

Step1. Construct difference matrix D in order to select most suitable or relevant text from the T. the D is calculated by having difference of index of last matched symbol of S with current index of match symbol in T iteratively except for the first symbol of S as in case of first symbol, the last symbol index is 0.

Step2. Calculate vectors R and E using following equations:

$$R=D \bmod 26 \quad (1)$$

$$E=D/26 \quad (2)$$

Step3. Estimate the maximum dual pattern repetition in R and store in a column matrix P. Now, select the largest row of P and denote it as P_{max}. The corresponding rows of R and E vectors are also selected and put in R* and E* vectors. The text from T corresponding to P_{max} is also chosen and put in T* as it is the most suitable text.

Step4. Apply BWT transform for rearranging the elements of R* in runs of similar elements.

Step5. Apply MTF encoding to further increase the correlation among the elements of R* after step 4.

Step6. Apply LZW compression technique to compress the R* as follows:

Construct the initial LZW dictionary using the integers between 1 and 26.

Update the LZW dictionary for every met symbol or symbol string. The concerning symbol or symbol string is encoded using the corresponding index in the dictionary.

Step7. Represent each element of R in binary form and concatenate them in order to obtain bit stream.

Step8. The bit stream is partitioned into groups of 14 bits, in each group, the first 9 bits are called G_1, next 3 bits are called G_2 and remaining bits are called G_3. The quotient and remainder of decimal representation of G_1 with respect to 26 are known as x and y respectively and decimal representation of G_2 is known as z.

Step9. Choose email addresses from K_1 by employing Latin square (as shown in Fig. 1) on x and y. select extensions of email addresses using z from A.

Step10. Alter chosen email addresses to incorporate random elements according to G_3 bits as follows

If the secret data bits are 01 then append a random symbol before '@' symbol.

Else if the secret data bits are 10 then append two random symbols before '@' symbol.

Else if the secret data bits are 11 then append three random symbols before '@' symbol.

Step11. Modify resultant email addresses (after step 10) in order to complete construction of K_2 set by using E*.

Step12. Construct stego-cover using T* as cover text and K_2 set as email addresses both.

3.2. Extraction phase

Step1. Get the stego-cover. Extract numeric elements of K_2 before '@' symbol to construct the vector E. If there is not any numeric element then E will be 0.

Step2. Count the number of characters before numeric values and if characters are 4 then secret data bits are 00, or 5 then secret data bits are 01, or 6 then secret data bits are 10, otherwise the secret data bits 11. Store these bits in G_3.

Step3. Extract first two elements from K_2 to obtain x and y by employing Latin Square of Fig. 1. Also extract email address extension to obtain z. Now, calculate G_1 and G_2 for each group of 12 bits by using the following equations:

$$G_1 = (x \bullet 26+y)_2 \quad (3)$$

$$G_2 = (z)_2 \quad (4)$$

Step4. Concatenate G_1, G_2 and G_3 in same order to obtain compressed bit stream.

Step5. Decompress the bit stream using LZW decoding:

Construct the initial LZW dictionary using the integers between 1 and 26.

Update the LZW dictionary for every met symbol or symbol string. The concerning symbol or symbol string is encoded using the corresponding index in the dictionary.

Step6. Apply MTF decoding to obtain R* (of step 4 of embedding phase).

Step7. Apply BWT to obtain original R* (of step 3 of embedding phase).

Step8. Estimate original difference D using R* and E as follows:

$$D = R + (26 \bullet E) \quad (5)$$

Step9. By using elements of D, extract the elements of S through T*, in the stego cover.

Thus, the secret data is obtained at receiver side. But, the receiver must have A and K_1 beforehand to extract the secret data.

IV. EXPERIMENTAL RESULTS

To analyze the performance of the text steganography methods, the hiding capacity is a prime parameter. In this section, we calculate these parameters for our proposed scheme and compare them with that of the recently developed data hiding schemes. Bit rate or hiding capacity is defined as the size of the hidden message relative to the size of the cover [16]. In this case, we can formulate bit rate as follows:

$$\text{Capacity} = \frac{\text{bits of secret message}}{\text{bits of stego cover}} \quad (6)$$

The novelty of our scheme lies into the combination of compression method which is BWT + MTF + LZW coding and also the way number of elements of email addresses are used to refer to secret data bits. We also add randomness to the selected email addresses while choosing number of elements in the email addresses for hiding the secret data bits which further improve the security aspect of the algorithm. Our proposed scheme is implemented in MATLAB running on the Intel® Core 2 Duo 2.20 GHz CPU, and 3GB RAM hardware platform. The secret message used in our experiment is "behind using a cover text …….. the intended recipient."(just for an illustrative purpose) and the cover text is "in the research area of text steganography,……at least 16 bits." (just for an illustrative purpose). Here, '…' is sign of the continuation of the text. The complete cover text and secret data can be referred from [16]. The secret message has 200 characters with spaces and without quotation marks. The cover text T* has the 847 characters without spaces and without quotation marks. The stego cover consists of the chosen cover text shown in Fig. 1 and the chosen and modified email address (K2). The employed Latin square is shown in Fig 1. According to Eq. (6), capacity has been computed as 7.03% for this example. The performance of our method is increased because the combination BWT and MTF increases the correlation among the data [17]. So, the LZW compression algorithm gives better compression ratio hence the hiding capacity is increased.

V. CONCLUSIONS

In this paper, we have proposed an email based text steganography method. The method makes use of the combination of BWT + MTF + LZW codlings to achieve higher capacity. It also uses number of elements of email addresses to refer to the bits of the secret data. To increase the security of the proposed method, the random elements are added into the email addresses to enhance randomness. In this method, the forward mail platform is used to hide the secret data. Our scheme performs better than the existing state of the art schemes like [16] in terms of hiding capacity.

Table I: Comparison of hiding capacity

Method	Capacity (%)	Explanation
Mimic functions [4]	1.27	Computed using given secret message at http://www.spamimc.com
Winstein [6]	0.5	Based on the referred paper
Sun et al.'s L-R scheme [10]	2.17	Computed using the given sample in Wang et al. (2009a)
Wang et al. [11]	3.53	Computed using the given sample in Wang et al. (2009a)
Listega [9]	3.87	Based on the referred papers
Satir and Isik [16]	6.92	Based on the given example of the same article
Proposed Method	**7.03**	**Calculated by employing the same example of [17]**

Rows	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
2	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
3	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
4	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
5	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
6	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
7	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
8	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
9	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
10	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
11	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
12	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
13	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
14	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
15	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
16	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
17	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
18	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
19	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
20	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
21	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
22	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
23	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
24	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
25	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
26	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

Fig. 1: Latin Square

References

[1] A. Gutub and M. Fattani, "A Novel Arabic Text Steganography Method Using Letter Points and Extensions," WASET International Conference on Computer, Information and Systems Science and Engineering (ICCISSE), Vienna, Austria, pp: 28-31, May 25-27, 2007.

[2] J. Y. Liang, C. S. Chen, C. H. Huang, and L. Liu., "Lossless compression of medical images using Hilbert space-filling curves," Computerized Medical Imaging and Graphics, vol. 32(3), pp. 174-182, 2008.

[3] P. Wayner, "Mimic Functions," Cryptologia vol. 16(3), pp. 193-214, 1992.

[4] P. Wayner, "Disappearing Cryptography" AP Professional, Chestnut Hill, MA (1996)

[5] K. Maher, TEXTO. 1995. ftp://ftp.funet.fi/pub/crypt/steganography/texto.tar.gz.

[6] K. Winstein, "Lexical steganography through adaptive modulation of the word choice hash", Secondary education at the Illinois Mathematics and Science Academy, January 1999.

[7] H. Nakagawa, K. Sanpei, T. Matsumoto, T. Kashiwagi, S. Kawaguchi, K. Makino and I. Murase, "Meaning Preserving Information Hiding _Japanese text Case," IPSJ Journal, Vol.42, No.9, pp. 2339 - 2350, 2001. (In Japanese)

[8] B. Murphy and C. Vogel, "The syntax of concealment: reliable methods for plain text information hiding," In Proceedings of the SPIE Conference on Security, Steganography, and Watermarking of Multimedia Contents, San Jose, CA, vol. 65(05), 2007.

[9] A. Desoky, "Listega: List-Based Steganography Methodology," International Journal of Information Security, Springer-Verlag, vol. 8, pp. 247-261, April 2009.

[10] X. Sun, G. Luo , and H. Huang, "Component-based digital watermarking of Chinese texts," Proceedings of the 3rd international conference on Information security, Shanghai, China, 2004.

[11] Z. H. Wang, C. C. Chang, C. C. Lin and M. C. Li, "A reversible information hiding scheme using left-right and up-down Chinese character representation," Journal of Systems and Software, vol.82, no.8, pp.1362-1369, 2009.

[12] Z.H. Wang, C.C. Chang, T.D. Kieu, and M.C. Li, "Emoticon-based text steganography in chat," In: Proceedings of 2009 Asia-Pacific Conference on Computational Intelligence and Industrial Applications vol. 2, Wuhan, China, pp. 457–460, 2009.

[13] R. Stutsman, C. Grothoff, M. Attallah, and K. Grothoff, "Lost in just the translation," in Proc. ACM Symp. Applied Computing, pp. 338–345, 2006.

[14] L-Y. Por, T.F. Ang, B. Delina, "Whitesteg: a new scheme in information hiding using text steganography" WSEAS Transaction on Computers, Vol. 7, pp. 735–745, 2008.

[15] L-Y. Por, K-S. Wong, and K-O. Chee, "UniSpaCh: A Textbased Data Hiding Method Using Unicode Space Characters," Journal of Systems and Software, vol. 85, no. 5, pp. 1075-1082, 2012

[16] E. Satir, and H. Isik, "A compression-based text steganography method," Journal of Systems and Software, vol. 85(10), pp. 2385-2394, October, 2012.

[17] L. Bin, N. Guiqiang, L. Jianxin, and Z. Xue, "BWT-based Data Preprocessing for LZW," International Conference on Multimedia and Signal Processing (CMSP), 2011.