International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE

Advanced Technology and Science

Original Research Paper

New Approach in E-mail Based Text Steganography

Kemal Tutuncu *[1], Abdikarim Abi Hassan

Accepted 26th May 2015

Abstract: In this study combination of lossless compression techniques and Vigenere cipher was used in the e-mail based text steganography. It makes use of email addresses to be the keys to embed/to extract the secret message into/from the email text (cover text). After selecting the cover text that has highest repetition pattern regarding to the secret message the distance matrix was formed. The members of distance matrix were compressed by following lossless compression algorithms as in written sequence; Run Length Encoding (RLE) + Burrows Wheeler Transform (BWT) + Move to Front (MTF) + Run Length Encoding (RLE) + Arithmetic Encoding (ARI). Later on Latin Square was used to form stego key 1and then Vigenere cipher was used to increase complexity of extracting stego key 1. Final step was to choose e-mail addresses by using stego key 1 (K1) and stego key 2 (K2) to embed secret message into forward e-mail platform. The experimental results showed that proposed method has reasonable performance in terms of capacity and also higher security in terms of complexity.

Keywords: *Text steganography, Latin square, Vigenere chipher, Stego key, BWT, MTF, RLE, ARI*

1. Introduction

The case of safety and security of information and especially secret communications has led to the introduction of several methods for secret message. Among these methods steganography is a rather new method. Steganography, the art of information hiding, has been used around for thousands of years, with the earliest examples coming from as early as 440 B.C.

Steganography is the traditional method used in information hiding. The use of systems like cryptography it makes the output of these methods very conspicuous. The main reason behind this is that the systems make messages unreadable to anyone except the intended recipient and does not hid its existence. The design of steganographic approaches is aimed at hiding the message or (data) and ensuring that its existence is hidden from observers.

Modern steganography can be applied to text, images, audio and video as shown in Fig.1. However, text steganography, has received less interest recent years, primarily due to the difficulty in finding redundant bits in text files and the lower capacity to hide information than the other mediums. This should not be the issue, as text Steganography has many advantages over the other mediums which make it model for effective Steganography.

One advantage of text Steganography over images and audio is that while they are both susceptible to compression due to their use of redundant data, this is not an issue with text Steganography as even though text contains redundancy, it cannot be removed or compressed. In addition the advantage to prefer text Steganography over images and audio is its smaller memory occupation and simpler communication. Text is also still one of the major forms of communication in the world, both in digital and printed form, and there are not many people who do not have access to text.

Text Steganography can be classified into three main groups: Structural, random and statistical generational, and finally linguistic. The main characteristics of structural text Steganography is that it modifies the physical form of the text, for instance through the appending white spaces and linespaces. On the other hand, random and statistical generation entails providing the cover text. This can be in a random manner or depending on specified input. While linguistic Steganography uses contents of natural language, for example verbs, nouns, adjectives and so on. Linguistic Steganography are classified into two: syntactic and semantic Syntactic text Steganography deals with changing the format of the text without considerable alteration to the meaning or tone.

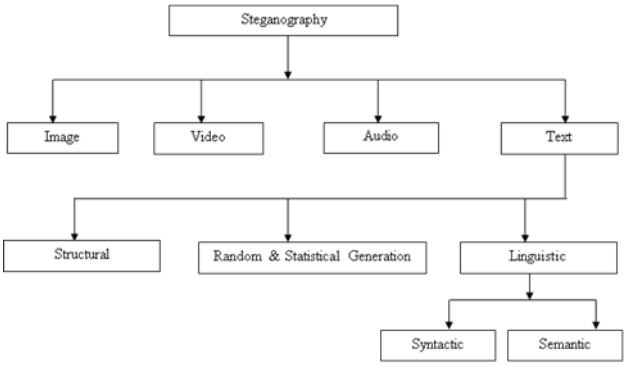

Figure 1. Categories of steganography

In this study, an email based data hiding method which makes use of combinatorial lossless compression to increase the hiding capacity and Vigenere cipher to increase the complexity was proposed. Run Length Encoding (RLE) + Burrows Wheeler Transform (BWT) + Move to Forward (MTF) + Run Length Encoding + Arithmetic Encoding (ARI) algorithms were used to have better compression ratio. Additionally Vigenere cipher was used to increase complexity of the system. The secret data was embedded into the email ids of forward mail platform. The most appropriate cover text is chosen from the text base by using distance matrix. After the stego cover determination we make use of distance matrix to find out email ids of a forward mail platform. Then e-mails were chosen from the previously arranged email address list. This email address list is used as a global stego key that is shared between both the sender and the receiver in advance. Capacity metric of the proposed method was evaluated by using obtained e-mail addresses.

The rest of the paper is structured as follows. Section II provides the related works. Section III and Section IV describe the proposed method and the experimental result. Finally, the

[1] *Faculty of Technology, Selcuk University Campus,Konya,Turkey*
** Corresponding Author: Email: ktutuncu@selcuk.edu.tr*

conclusion is drawn in Section V.

2. Related Works

In [1], [2], Satir and Isik used email to be the cover of the secret data. Sending same e-mail to many recipients were the key idea of their studies. There may not be a proper method to validate whether the email address of each recipient is valid but sending an email to the respective address. In those papers [1], [2], email addresses are used to hide the stego keys, regardless they are valid addresses or not [3]. In [3] the authors proposed improvement of the study [2], such as the randomity of the generated email addresses. They claimed that random generated email addresses may increase the security level. In [4], the authors proposed combination of BWT + MTF + LZW coding algorithm to increase the hiding capacity and adding some random characters just before the '@' symbol of email ids to increase the randomness.

In a chat application, Wang et al. [5] employ emoticons (emotional icons) to hide a secret message. This is intended to improve the performance of other chat-based steganography. The proposed method is developed based on the assumption that the use of emoticons in chatting is high, where many recent applications using it. It is claimed that this method is able to raise up the capacity of the secret data to be embedded as well as to provide an easy to use application. This method requires both sender and receiver have an exactly same emoticon table containing some emoticon sets. In this case, the order of the emoticon affects the meaning of the secret message. Therefore, there must be a synchronization process before the communication begins. Practically, this method requires the users to input many emoticons in order to send the secret data. This may have an effect on the user convenience [4]. Wayner [6,7] 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. Other researches have been carried out which employ varied communication medium, such as in [8], [9]. Hiding the secret in the text is also investigated by Desoky [10]. He proposes Listega, a method to hide data which exploits the popularity of list of items, such as song, food, drink and car. Another text steganography given by Maher [11] 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 [12], another important method is synonyms-based approach [13-15]. Unlike [11], the method uses legitimate words and sentences having appropriate preciseness. Thus, the visual attack will have wee significance on these types of methods.

3. Proposed Method

3.1. Embedding Phase

Let, S: secret message T: A text base of cover texts in which the secret message will be embedded. K1: 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 as in [1,2,3,4]. **Step2.** Calculate vectors R and E using following equations: R=D mod 26 E=D/26

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 Pmax. The corresponding rows of R and E vectors are also selected and put in R* and E* vectors. The text from T corresponding to Pmax is also chosen and put in T* as it is the most suitable text.

Step4. Apply Run Length Encoding (RLE) + Burrows Wheeler Transform (BWT) + Move to Forward (MTF) + Run Length Encoding + Arithmetic Encoding (AE) in the written order for compressing elements of R*

Step5. Represent each element of R in binary form and concatenate them in order to obtain bit stream.

Step6. The bit stream is partitioned into groups of 12 bits, in each group, the first 9 bits are called G1, next 3 bits are called G2. The quotient and remainder of decimal representation of G1 with respect to 26 are known as x and y respectively and decimal representation of G2 is known as z.

Step7. Choose two characters as K1 by employing Latin square on x and y. Select extensions of email addresses using z from A.

Step8. Use Vigenere cipher to alter two characters obtained from Latin Square. Thus these two new characters will be used to choose e-mail address. Choose email addresses that starts with modified K1 before @ and ends with z after @.

Step9. Modify resultant email addresses in order to complete construction of K2 set by using E*. Combine modified K1 and K2 (obtained from z) to form e-mail addresses.

3.2. Extracting Phase

Before starting extraction phase both the sender and receiver must have email address list, three bits corresponding of email extension, Vigenere cipher key, RLE, MFT, BWT and ARI tables in advance.

Step1. Get the stego-cover. Extract numeric elements of K2 before '@' symbol to construct the vector E. If there is not any numeric element then E will be 0.

Step2. Extract first two elements from K2 to obtain Vigenere values of x and y by employing Vigenere Cipher.

Step 3. Extract the values of x and y by employing Latin Square Also extract email address extension to obtain z.

Step 4. Calculate G1 and G2 for each group of 12 bits by using the following equations: G1 = (x • 26+y) G2 = (z)

Step4. Concatenate G1 and G2 in same order to obtain compressed bit stream.

Step5. Decompress the bit stream using Arithmetic Encoding (AE) + Run Length Encoding (RLE) + Move to Forward (MTF) + Burrows Wheeler Transform (BWT) + Run Length decoding to obtain original R*

Step7. Estimate original difference D using R* and E as follows: D = R + (26 • E)

Step8. By using elements of D, extract the elements of S through T*, in the stego cover.

4. Results

The proposed method was prepared by programming in MATLAB. The computer used for running the program was Intel Pentium7, with 8 GB RAM. Later on same computer was used to analyse performance of proposed method or program. The performances of the text steganography methods can be analysed by capacity ratio or complexity. In this study both parameters are investigated. In order to calculate the capacity ratio secret data whose length is 196 characters is used. The secret message is as follows:

'behind using a cover text is to hide the presence of secret

messages the presence of embedded messages in the resulting stego text cannot be easily discovered by anyone except the intended recipient'. Cover data set was ranging from 700 characters to 4000 characters. 35 cover data set was used. It has been seen that the success of the embedding is directly related with length of the secret message and cover text. The smaller size the cover text the less successful the embedding is. Two factors are important here; Firstly the length of cover to possibly include all the characters of the secret message due to its length and secondly the match between the content of the cover message with the frequencies off English alphabet. If a cover text is formed according to the frequency of English alphabet it will cause better capacity ratio for text steganography.

Table 1. Applying text message (200 char) to different cover text

Cover	Cover Length (char)	Result	Number of e-mail addresses
CT1	300	Fail	-
CT2	800	Fail	-
CT3	1500	Fail	-
CT4	3993	Success	1957

As can be seen from Table 1. CT1, CT2 and CT3 cover texts resulted in not hiding secret message. No e-mail address is produced to hide secret text. CT4 successfully hided the secret text. Here the key point is how much repetition occurs when we created distance matrix. If we have much more repeated number in distance matrix then the compression ratio of the combined lossless compression techniques will increase. This will cause increase in capacity ratio, thus decrease number of e-mail addresses. From this point of view it is quite important to choose cover text to increase the capacity ratio in text steganography. Briefly explaining not only the length of the cover data determines the successfulness of the embedding process, but also variation of the characters. Longer the cover data actually increases the possibility of providing more varied characters. Therefore, successfulness of the embedding process depends on the appropriate selection of the cover data [3].

Complexity is one of the important aspects of the steganography. It is directly related with extraction of hidden information by an adversary in an easy way or not. In this study compression algorithms and two combinatory based coding (Latin square and Vigenere cipher) were used to increase the complexity. Complexity analysis of the proposed system will be explained. Before explanation one must keep in mind that global key and the key of Vigenere cipher were shared between sender and receiver in advance. Thus, if any adversary would like to break the system he will not have these two keys. The step that an adversary will follow to break the system as follows:

1. Adversary must extract compressed R*. In order to implement this action he/she has to obtain G1from x, y and G2 from z. At his point x and y can only be obtained by using the Vigenere cipher that consists of two characters. Finding correct combination of these two characters will require 26*26 combinations. Finding correct combination will let adversary to find the Latin Square corresponding of x and y. Afterwards x and y can be obtained from Latin Square. The z can be obtained by the extension of e-mail address. Adversary has to calculate 8 combination per each e-mail. Thus, if the number of e-mail address is 8 then 8^N combination must be calculated. Later on each z is expressed in binary form and then combined with x and y to obtain G1 and G2. In order to find out correct bit stream R*, 8^N*26^2 combination must be formed.

2. R is obtained by solving R*. As it is known that D = R + (26 • E) and each e-mail address is checked to see whether it has number before @ sign or not. In case of including numbers before @ sign these numbers are used for obtaining members of E. These numbers can belong to e-mail address naturally or added to e-mail address to hide members of E during embedding phase. If we express number of numbers in an e-mail address as m then the adversary will implement 2^m combination to extract the members of E.

3. Obtained each R and E must be tested to each other to see the correct hidden message.

Thus complexity of the proposed system is

$Comp\_proposed\_system = 26^2\ 8^N \times \prod_1^N 2^m$

Complexity of the system is calculated without regarding the compression algorithms that are used to have combination of RLE + BWT + MTF + RLE + AE. In the study of [2] the authors didn't make use of complexity of obtaining decompressed text. In order to make comparison with this study only the complexities contribution of k1, k2 and Vigenere cipher were used. In [2] complexity was followed as $8^N \times \prod_1^N 2^m$. Thus proposed method increased the complexity as the amount of 26^2 .

5. Conclusion

Proposed method is extension of text-based steganography that makes use of email environment to hide the secret text. Capacity ratio and complexity of the proposed method are improvements of the previous e-mail environment based text steganography methods [1-4] in terms of complexity. Reasonable capacity was obtained. By using combination of Run Length Encoding (RLE) + Burrows Wheeler Transform (BWT) + Move to Front (MTF) + Run Length Encoding (RLE) + Arithmetic Encoding (ARI) algorithms the size of the secret message was reduced. Vigenere cipher added extra complexity or security to the system for obtaining stego key (K1). The study also showed that the length and the content (variation) of the cover text directly related with the success of the embedding process. Short cover text with not well distributed alphabet content (doesn't match with the frequency of the English alphabet) will result failure in embedding process.

Other combination of lossless compression techniques can be tried for increasing the capacity ratio of the email environment based steganography. Additionally some other techniques that put randomness to the email addresses can be used to increase the complexity of the related system.

References

[1] E. Satir and H. Isik, A compression-based text steganography method, The Journal of Systems and Software Science Direct, vol. 85, issue 10, pp. 2385-2394, 2012.

[2] E. Satir and H. Isik, A Huffman Compression based Text Steganography Method, Multimedia Tools Appl, September 2012.

[3] Tohari Ahmad, Melvin S. Z. Marbun, Hudan Studiawan, Waskitho Wibisono, and Royyana M. Ijtihadie, A Novel Random Email-Based Steganography International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 4, No. 2, April 2014

[4] Rajeev Kumar , Satish Chand , and Samayveer Singh, An

Email based high capacity text steganography scheme using combinatorial compression, 25th International Conference-Confluence The Next Generation Information Technology Summit (Confluence), 25-26 September 2014, India

[5] Z. Wang, T. Kieu, C. Chang, and M. Li, Emoticon-based text steganography, in Proc. 2009 Asia-Pacific Conference on Computational, Wuhan, China, 2009.

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

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

[8] L. Por, K. Wong, and K. Chee, UniSpaCh: a text based data hiding method using unicode space characters, Journal of Systems and Software, vol. 85, issue 5, pp. 1075-1082, 2012.

[9] M. Topkara, U. Topkara, and M. Atallah, Information hiding through errors: a confusing approach, in Proc. SPIE International Conference on Security,Steganography, and Watermarking of Multimedia Contents, San Jose, CA, 2007

[10] A. Desoky, Listega: list-based steganography methodology, International Journal of Information Security, vol. 8, no. 4, pp. 247-261, 2009.

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

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

[13] 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)

[14] 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.

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

