A Secure Text Steganography Based on Synonym Substitution

Cao Qi[2], Sun Xingming[1,2] , Xiang Lingyun[2]
[1]Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology
[2]School of Computer and Communication, Hunan University, Hunan, 410082
sunnudt@163.com, cao_qi0104@163.com

Abstract—**In text steganography based on synonym substitution, the synonym number in synonym set is not always some powers of 2. In traditional methods, the redundant synonyms are always abandoned. Two different coding schedules, which made full use of the abandoned synonym to improve the security of the steganographic system, are proposed in this paper. The synonyms in a synonym set are dynamically divided into coded words and backup words. One method divides the cover text into blocks to hide the secret information for the first time and then uses a highly efficient coding schedule to embed the secret information for the second and third time. The other method employs the backup words to prevent the increase the number of synonym pairs, keeping the statistical characteristics of synonym pairs unchanged.**

***Keywords**- text steganography; text steganalysis; synonym substitution; synonym pairs;*

I. INTRODUCTION

The current text steganography is mainly based on format, invisible character and natural language[2]. When facing attacker's re-editing the latter is much safer and the easiest way to implement is steganography based on synonym substitution[3]. However, the synonym number in synonym set is not always some powers of 2, the abundant words are generally abandoned in coding.

In this paper, we propose two different methods to take advantage of the abundant synonyms. In one method, we divide the cover text into blocks, embed secret message in every block, the first time with method of traditional synonym substitution and the second and the third time a highly efficient coding schedule .To extract secret message exactly, the abandoned synonym is used to record the embedding process. In the other method, we dynamically divide the synonyms in a synonym set into coded words and backup words to prevent the increase of the synonym pairs number. Therefore, the statistical characteristics of synonym pairs were kept unchanged. Experimental results showed robustness of this steganographic method when attacked by steganalysis using the feature of synonym pairs.

II. REPETITIVE EMBEDDING SCHEDULE

A. Highly Efficient Coding Schedule

Decreasing the modification to keep the original features of cover is a good measure to resist all kinds of steganalysis. Literature[4]proposed a way that a modification of 1 bit at most can mean embedding of $[\log_2(n+1)]$ bits secret message in binary data stream. For a binary data stream $C=(0,1,0,1,1,0,1)$, its 7 revisable positions and itself mean totally 8 states. Therefore, secret message of 3 bits can be embedded when modify 1 bit or not in C. For example, if original cover C is $(0,1,0,1,1,0,1)$, embedded cover C' is $(0,1,0,1,1,0,1)$, $(0,1,1,1,1,0,1)$ or $(0,1,0,1,1,0,0)$,secret message " 000 ", " 011 ", " 111 " can be extracted respectively from C' by comparing with C.

B. Repetitive Embedding

If a receiver wants to extract secret message from repetitive embedding cover, s/he needs to recover the original cover, so the sender must use an extra $\log_2(n+1)$ bits to mark the modified position and send them together with the secret message, which is useless to enlarge the steganographic capacity. We use the abandoned synonyms to record the modified position. To descript accurately , symbols are defined as follows:

Definition 1 According to the order of synonyms appearing in cover C, we get a sequence D_n : $d_1, d_2, ..., d_n$.

Definition 2 The synonym set that Synonym d_i belonges to is named as A_j , j shows position of A_j in synonym database, a_{jk} means the k th synonym in A_j ;

Definition 3 N_i is the synonym number that A_i includes. $2^t \prec N_i \prec 2^{t+1}$, $(t=1,2,3,...)$;

For simplicity, set $N_i=3$. In order to extract data exactly for the receiver we use two words, each word has 3 synonyms, that is $3\times3=9$ states ($S_1, S_2 \cdots, S_9$) to mark embedding process. The embedding algorithm description is as following:

Step1 : Get D_n from original cover text and divide them into several D_9 . Encode the synonyms in the synonym database with the traditional schedule, in which the first two words were coded in every synonym set. Transform the secret message into binary data stream;

Step2 Choose a D_9 that is not managed yet. According to the secret message replace d_1 or not in first 7 positions of D_9 with traditional schedule, if bit that d_1 carried is not as same as secret data, replace d_1 ,else keep d_1 unchanged. d_8,d_9 can show 9 states($s_1,s_2...,s_9$) to mark embedding process;

Step3 Get the following 6 bits secret message, if the first three bits and last three bits are same, go to step4, else go to step5;

Step4 If both of two parts are not "000", calculate and get their decimal number i,

If $d_i ==$ a$_{j1}$,use the first state s_1 to mark .

Else use the second state s_2 to mark.

Replace d_i with a$_{j3}$

Step5 If the first part is "000", calculate the last part and get its decimal number i,

If $d_i ==$ a$_{j1}$,use the third state s_3 to mark .

Else use the fourth state s_4 to mark.

Else If the last part is "000", calculate the first part and get its decimal number n,

If $d_i ==$ a$_{j1}$,use S_5 to mark .

Else use the sixth state to mark..

Else calculate the two parts and get their decimal numbers i and j . According to $d_i == a_{i1}$ or not , $d_j == a_{j1}$ or not , $i \succ j$ or not, use 8 different states to mark. And replace d_i with a_{i3} ,replace d_j with a_{j3} (a_{i1} and a_{i3} shows the first and third position respectively in A_i that d_i belongs to. a_{j1} and a_{j3} shows the first and third position respectively in A_j that d_j belongs to) ;

Step6 Go to step2 until secret message is over.

For example, if original D_9 is: $a_1b_3e_2c_1c_3b_2a_3e_2c_2$, secret message is 1001011111010 .There are 4 synonym sets: $A_1:\{a_1,a_2,a_3\}$, $A_2:\{b_1,b_2,b_3\}$, $A_3:\{c_1,c_2,c_3\}$, $A_4:\{e_1,e_2,e_3\}$ in synonym database. After first embedding, D_9 is $a_2b_1e_1c_2c_1b_2a_2e_2c_2$. The decimal numbers of following two 3-bit are 7 and 2, according to embedding algorithm the fifth state is used to mark , that is e_2c_2 . Replace the 7th and 2nd of D_9 with a_3 and b_3 . Finally D_9 is $a_2b_3e_1c_2c_1b_2a_3e_2c_2$.

Secret data can be extracted exactly according to embedding algorithm.

C. Corresponding Analysis of Steganographic Capacity

If y means capacity of steganography system , n means the number of revisable position in cover text. y is a function of n, $y = f(n)$. When $N_i = 3$,for traditional schedule, one revisable position means that 1bit secret message can be embedded, so $y = n$;For repetitive embedding schedule,7 bits are embedded at the first time and 6 bits are embedded at the second and the third time in D_9 .So, we get $y = \frac{n}{9}(6+7) = \frac{13}{9}n$ .Comparatively , we improve the steganographic capacity by 44%.

III. KEEPING THE ORIGINAL FEATURE OF SYNONYM PAIRS

A. Feature of Synonym Pairs

As defined in[5], list all the words with same Semantic meaning in text C, we get a sequence with length l : $w_1,w_2,...,w_l$, if $l \geq 3$,get w_0 and w_{l+1} ,($w_0 \neq w_1$, $w_{l+1} \neq w_l$)to make a new sequence $w_0,w_1,w_2,...,w_l,w_{l+1}$,if there is $w_{i-1} \neq w_i = w_{i+1} \neq w_{i+2}$ ($i \in [1,l-1]$),we say there is synonym pairs about w_i . [5] proposed that the number of synonym pairs will increase after steganography based on synonym substitution with traditional coding schedule. It is easy to be attacked by steganalysis using the feature of synonym pairs. We propose a method that use the abandoned synonym to avoid the increasing of synonym pairs, keeping the original feature of cover text unchanged.

B. Embedding Algorithm

To descript accurately , symbols are defined as follows:

Definition 4 if words in synonym sets are coded currently, we call them current coding list(CCL);

Definition 5 If words in synonym sets which is not in CCL,we call it backup word(BW).

Definition 6 $mark[j][i]$ is 0 if the i th word in j th set is not in CCL,else $mark[j][i]$ is 1. frontbin[j] records the embedded data that the j th set carried last time, initialized as -1. lastcode[j] records the location of replaced word in synonym set, initialized as -1.

Transform the secret message into binary data, and the embedding algorithm description is as following:

Step1 search synonym that synonym database includes, find the position j of its corresponding set in database.

If frontbin[j]=-1,according secret message replace the synonym or not, change frontbin[j] as current secret message,change lastcode[j]. Go to step3;

Else go to step2;

step2 compare frontbin[j] with current secret data, If equal，replace the synonym with BW, get the BW's location pos in set,set mark[j][pos]=1,mark[j][lastcode[j]]=0;

Else according to secret message replace the synonym or not,change lastcode[j] and frontbin[j].

Step3 ,go to step1 until secret message is over.

Secret data can be extracted exactly according to embedding algorithm.

C. *Experimental results and comparison*

In order to evaluate our proposed schedule, some numerical experiments are performed. 100 experimental Chinese texts are downloaded on internet. We use traditional method and our proposed method to hide information. Fig.1 shows synonym pairs number of original text and traditional method. Fig.2 shows synonym pairs number of traditional method and our proposed method after embedding.

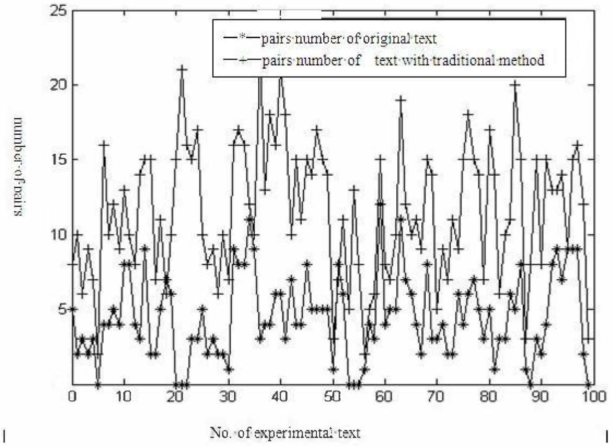

Figure 1. synonym pairs number of original text and traditional method

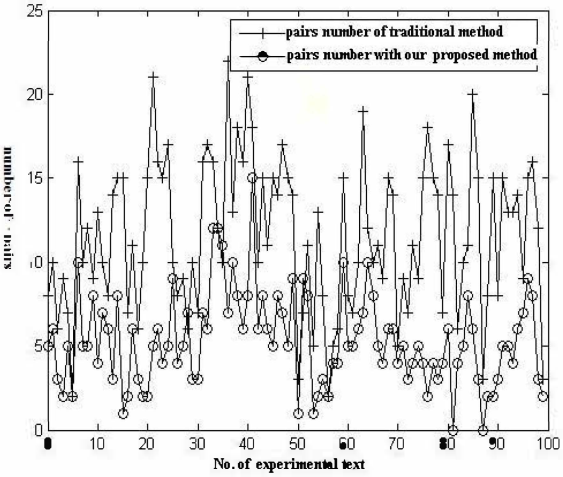

Figure 2. synonym pairs number of traditional method and our proposed method after embedding

As shown in Fig.1,pairs number of traditional method is obviously more than original text， but evidently, pairs number of our proposed method is less than traditional method While keeping the features of original text at utmost. We use the algorithm in [5] to detect steganographic text with traditional method and our proposed method, and set limen $\delta = 5\%$,the experimental result is as shown as Table 1.

TABLE I. DETECTIVE RESULT TO DIFFERENT EMBEDDING

Text type to be detected	Number of text	Number of false Positive	Probability of false Positive
traditional method	100	11	11%
our proposed method	100	71	71%

IV. CONCLUSION

Two novel method are proposed to take advantage of the abandoned synonym in traditional steganography based on synonym substitution. The first method improves the steganographic capacity by 44% than tradional method and the second method can avoid the steganalysis using the feature derived from synonym pairs.

ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (No. 60736016, No. 60973128, No. 60973113, No. 61073191 and No. 61070196), and in part by National Basic Research Program of China (No. 2009CB326202, and No. 2010CB334706), PAPD.

REFERENCES

[1] Petitcolas F A P,Anderson R J,Kuhn M G .Information hiding-A survey. Proceedings of the IEEE:Special Issueon Protection of Multimedia Content, 1999,87(7): 1062-1078.

[2] Topkara M,Taskiran C M,Delp E J. Natual language watermarking[C]//Proceedings of SPIE San Jose,CA,USA,2006:60720A

[3] Chiang Y L,Chang L P,Hseh W T,etal Natural language watermarking using semantic substitution for chinese text[C]//IWDW 2003.Heidelberg. SpringerBerlin, 2004: 129-140.

[4] Tian Yuan, Cheng Yi-min, Wang Yi-xiao. A novel method of data hiding, [J]Acta Electronica Sinica 2004, 32(9) : 1444-1447.

[5] Gang Luo,Xingming Sun,Lingyun Xiang. Steganalysis on Synonym Substitution Steganography.Journal of Computer Research and Development 45(10) : 1696—1703, 2008