A Text Information Hiding Algorithm Based on Alternatives

Liu Gongshen, Ding Xiaoyun, Su Bo and Meng Kui
School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai China
{lgshen, leonard520, subo, mengkui}@sjtu.edu.cn

***Abstract*—In this paper, we analyze existing text-based steganography techniques, and propose an efficient information hiding algorithm for text based on substituted conception. In this algorithm, candidate units are replaced by similar conception, including synonym, homonym or special character. An improved encoding algorithm is also proposed which is not only enhances the robustness, but also enlarges the capacity of secret message. It is proved by the experiments that the proposed algorithm has good performances both in security and capacity.**

***Index Terms*—stego-text; encoding; replaceable unit; information hiding;**

I. Introduction

Information hiding means concealing the information itself and its location. The hidden message can be text, passwords, images, graphics or sound. It also can be the file in one's computer. At the meantime the information carrier can be a general digital image, digital video, software, digital audio, or text file. Research in information hiding has a long history and in recent years it becomes a research hotspot again. Judging from research papers published in this area, most researches focus on how to hide open information and digital watermark with image file. This is mainly due to that image file has large redundant information capacity, and image processing is more intuitive. Since the existence of redundant information, we can hide some information without the suspicion of the observer. But text file does not contain any redundant information for secret information transmission. Information hiding in text is not a simple mission.

It is generally considered that it is Mikhail J. and M. Attalla of Purdue University who first proposed the concept of natural language text information hiding [3] in 2000. Natural language information hiding technology takes advantage of natural language processing technology. It embeds secret information by changing the attributes of the original text while keeping their meaning. There are three kinds of text information hiding methods.

The first one is based on changing the format of the text. This method is for the text with a certain layout format or file structure. Compared to the plain text, the formatted text contains more redundant capacity in format information. For example, reference [1] hides information by adjusting the line spacing, word spacing, font, and character size, building signature or using special formats like document head. This method possesses favorable commonality. But it can't hide large information, and once the algorithm detail is published, it will be easy to be cracked.

The second one is based on the syntax of text. Usually, it changes the syntactic structure of sentences. The commonly used methods include moving the position of adjunction, adding formal subjects, changing the active/passive form, changing the statement sequence. Especially, Mimicry Steganography Formulation uses context-free method to describe the structure of the sentence to construct steganography text, and chooses different sentence structure to hide information. It has greatly helpful to promote syntax-based information hiding technology.

The third one is based on the semantic of the text. Many researchers are working on it now, especially on synonyms replacement. There are some good works on the consistency of the text and encoding of the synonyms of the text. Reference [2] works on the synonym replacement method. It uses the text content of blog and forum posts as the carrier, and proposes that we can use spelling errors to make word replacement and information hiding. However, the research about synonym-based method in China is far from enough. Reference [5] proposes an information hiding method based on punctuation replacement. The method deletes or adds the colon between two independent clauses (sentences) according to certain rules. Thus we can take advantage of this feature to express '0' by keep colon or change semicolon to colon and express '1' by keep semicolon or change colon to semicolon. Reference [4] does the word segmentation first and then synonyms replacement. But it doesn't take the context of the synonyms into consideration. A kind of new approach is constructed for the webpage information hiding [15]. The secret message is required for modification using forward transform technology before hiding. A kind of grouping 5-bit scheme and algorithm are proposed to convert the hiding data into some invisible characters. It can hide 10 bits in each end tag of page line. In Reference [16], a text steganography system for spelling languages is proposed based on Markov Chain source model and DES algorithm. It can work reliably with the capability of immunity from regular operations, such as formatting, compressing and sometimes manual altering operation in text size, front, color and the space between words. It's suitable for hiding short information in online communication, such

doi:10.4304/jsw.8.8.2072-2079

as E-mail, MSN Messenger, QQ, instant conversation and the short message on mobile phone.

In this article, we propose an information hiding method based on the substituted conception. In the second chapter, we show our text hiding system platform which combines the encoding of the secret information with searching of self-definition knowledge base. We introduce the definition of the substituted conception, knowledge base and coding algorithm first. Furthermore, we also give a detail description about how to use the conceptions. An improved coding algorithm--LZ code along with hamming code, is presented in Chapter II too, which makes the system more robust and capacity. Performance testing of the system platform is listed in the third chapter.

II. The information Hiding Method Based On The Substituted Conception

Substituted conception is defined as: for a certain part in the text, we can find something in our knowledge base and replace them with the same or similar text for representation. And the replacement will not change the meaning of the text and it also does not cause ambiguity or error.

The synonym is a king of substituted unit. There are a lot of researches methods of the substituted method based on synonyms in the foreign country and have a good result in consistency or the encoding of synonyms. However, for the Chinese own characteristics, Chinese information processing technology is far from mature. It needs to put in more effort in the text information hiding. We will present an integrated replaceable unit definition, Knowledge base generation and encoding algorithm of information hiding system platform.

The flow chart of the platform is as follows shows.

A. Substituted Conception

1) Use the synonyms as a Substituted Conception

The method based on the synonym substitution is the most widely used method in Chinese natural language information hiding method. In Synonym substitution algorithm, we select words that appears in our knowledge base and encode them by some certain encode method to embed information.

The so-called synonyms, generally defined as "in one language (Chinese here), synonyms shares the exactly meanings between two or more words in some or all the semantic". [6]. in the original Chinese synonyms replacement method, researchers did not do the word segmentation. They check in the vector text in order to find a synonym in the dictionary and do the synonyms replacement to embed information. This method has an obvious defect; it doesn’t do the word segmentation so when it finishes the synonymous substitutions, it is still very easy to change the meaning of the original semantics, resulting in the practical application of this algorithm is of little significance. Therefore, the general Chinese synonym replacement algorithm must first process an automatic segmentation.

We first convert the message we want to hide to the binary string. Then we do a segmentation of the whole text and go through every word of the whole text which finished the word segmentation to determine whether the word is in the dictionary. if the word is in our knowledge base, we first get the encoded information by the secret messages. Then we decide which words in this group should be substituted by the number of the group and the position of the words. The encoding method will be in-depth discussion in chapter 5, including the additional length check code, this can be very effective to improve the robustness to prevent the amount of information loss caused by destruction with hidden information file, do this things till all the information hiding finished and joined the logo identifies the end of bit. if we have finished traversing the text but still not get embedded work done, the information hiding process returned as failure one. Get hidden information algorithm, the first step is doing the word segmentation of the text which has the hidden information. After traversing each of its words, the user determines whether the word is in the thesaurus base. If it is in the thesaurus base, according to the number and location of the group of words in the word and the embedded algorithm coding and decoding to obtain the response of binary string, and repeat these steps until the end of the flag.

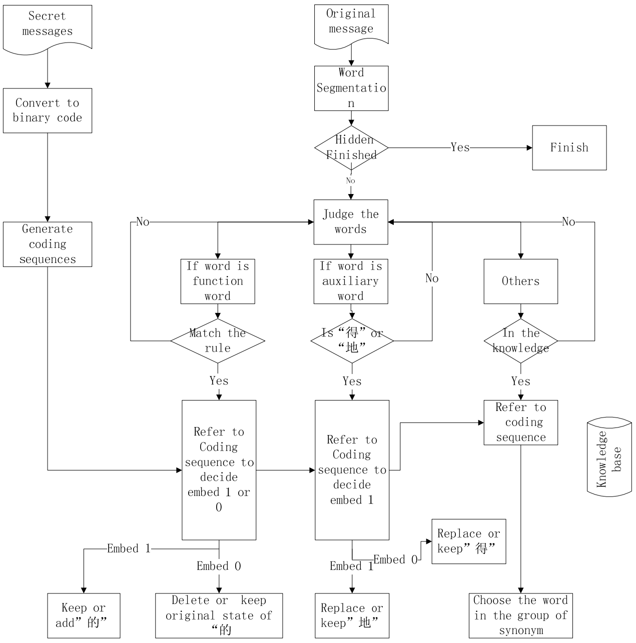

Figure 1. The diagram of information hiding algorithm by substituted conception

This article uses the "Hownet" and "The Dictionary of Synonym Words" as mentioned dictionary above because it has a large size of the lexicon and it has a collection of nearly 70,000 words, all the lexicon compiled by significance. It is such a dictionary that it is suitable for use as a synonym-based text to hide the technology. In fact, due to the use of flexibility of the Chinese, direct use of this lexicon will encounter the problem of inconsistent semantics before and after the replacement text. For example, the "fight" and "hit" are synonyms, but "taxi" must not be replaced by "hitting the car". Please refer to in-depth study on this issue in chapter 2.2

2) Use Function Words as Substituted Conception

Through the incomplete word frequency statistical study on different kind of text styles, we found that the function words such as ‘的’, ’了’, ‘是’ are of high frequency in Chinese. Even if the original text is a science and technology article, the weakness of other semantic based text steganography approach will not appear. When the text carrier contains a large amount of terminology or highly restrictive words, a slightly change on the word order or substitution with synonyms will change the meaning of the original sentence or even make it incomprehensible. In some circumstances, function words can be added or deleted without changing the original meaning or the quality of the text, which will not draw the attention of other detectors. The algorithm proposed in this section will take advantage of this feature to hide information.

We use the high frequency function word ‘的’ as the first object to study. Firstly, we summed several rules that can be embedded with hidden information. In these rules, the ‘的’ in the ‘的’ structure can be added or deleted without other people’s awareness.

rule 1 ‘的’ can be removed when using monosyllabic adjective as attribute

e.g.:新发现——新的发现

rule 2 nouns as attribute

e.g.:玻璃门——玻璃的门

rule 3 pronouns as attribute

e.g.:我祖国——我的祖国

rule 4 “其他”、“其它”、“其余”as attribute

e.g.:其他意见——其他的意见

We can conclude that all the function words that meet the rules above can be added or deleted.

In the concrete realization, firstly, convert the secret message into binary strings. Then traverse the text and apply word segmentation operation. Determine on each function word, if there existed ‘的’ that can be added or deleted, then decide to add or delete the ‘的’ according to the binary string of the secret message. Here, we make the rule that the added or originally existed ‘的’ represents 1, while the deleted or originally not existed ‘的’ represents 0. Repeat the steps above until finish the embedding work of secret message and then add the flag that represents the end of bit. If the embedding work is not finished after traversing all the words in text, then the embedding work is considered to be failed. When recovering the secret, split the sentence of the encrypted message. For each sentence, if there exists ‘的’ that can be added or deleted, then decide if there should be a ‘的’ in the text. If there should be, then add 1 to the recovering message, otherwise, add 0. Repeat the steps until the end flag..

3) Use Homophones as Substituted Conception

Other than a large amount of synonyms in Chinese, there also exists a large amount of homonym replacement such as standardized forms of words and non-standardized ones. The difference between the homophones and synonyms is that they are exactly the same in pronunciation and meanings but different in written form. We can also realize text steganography by replacing homophones between standardized forms of words and non-standardized ones.

In this section, the text steganography approach based on the substitution with homophones will be divided into two types.

(1) Standardized forms of words and non-standardized ones. Chinese language is profound. In its development process, there appears a large amount of standardized forms of words and non-standardized ones. Although some of them is rarely used and even eliminated, the rest are still in using and can be used in the technology proposed in this section." the First Series of Standardized Forms of Words with Non-Standardized Variant Form"[11] lists 338 groups of commonly used non-standardized ones and recommend ones, these words can be directly add into the homophones dictionary.

(2) Structural Particle. The homonym substitution words mentioned in (1) are of very low frequency in the text, so the content that can be embedded is strongly restricted. Therefore, we come to the idea about using the structural particles like ‘的’, ‘得’, ‘地’, which appears with high frequency in any text, to optimize the text steganography method based on homophones substitution. Since ‘的’ is already used is the previous section, to avoid confusion, we will discuss on the mutual substitution between ‘地’ and ‘得’.

‘地’ follows the adverbial when it is used as structural particle, representing a relationship of modification between adverbial and central word, which can be used for homonym substitution with ‘的’. One of the typical structure is (adverb, adjective) + ‘地’ + (verb, adjective).

‘得’ follows the verb or adjective when it is used as structural particle, con-nects the complement that represents the extent or results. Or, it is used be-tween verb and complement, representing possibility. It can also be used for homonym substitution with ‘的’. One of the typical structure is (verb, adjec-tive) + ‘得’ + (verb, adjective, adverb).

In the concrete realization, firstly, convert the secret message into binary strings. Then apply word segmentation operation upon the original text. Traverse the encrypted message, for each particle ‘的’ and ‘得’, reserve or replace them according to the binary string of the secret message. Here, we make the rule that ‘地’ represents 1, while ‘得’ represents 0. Repeat the steps above until finish the embedding work of secret message and then add the flag that represents the end of bit. If the embedding work is not finished after traversing all the words in text, then the embedding work is considered to be failed. When recovering the secret, split the sentence of the encrypted message. For each sentence, if there exists ‘的’ that can be added or deleted, then decide if there should be a ‘的’ in the text. If there should be, then add 1 to the recovering message, otherwise, add 0. Repeat the steps until the end flag.

B. Setup of Knowledge Base

As mentioned in 2.1, using "The Dictionary of Synonym Words" directly cannot meet the condition of large enough vocabulary. And it will also encounter with the problem of semantic inconsistency in the context. In this section, we proposed a method for the construction of a better knowledge base, mainly used in the algorithm of using synonym as replaceable unit, and the other part used in the algorithm of using homophones as replaceable unit.

1) "The Dictionary of Synonym Words"

On the basis of "The Dictionary of Synonym Words", the Harbin Institute of Technology Information Retrieval Laboratory made the "The Dictionary of Synonym Words modified version". The extended version referenced many electronic dictionary resources. It get the frequency of a word in People's Daily corpus and only retained the words with the frequency no lower than 3 (the statistical results of a small-scale corpus). At last, a total of 77434 words are included by the dictionary. The dictionary organized all the included words together in accordance with the tree hierarchy. The vocabulary is divided into three categories as large, medium and small. There are 12 large categories, 97 medium categories and 1428 small categories. Then divide the small categories into word groups. Words in each word group are further divided into several lines. Words in the same line holds the same meaning (or close meaning) or strong correlation.

2) HowNet[10]

The cause of semantic inconsistency in the context is that even synonymous will have the difference of exactly same and partly same. In the synonym dictionary proposed in 2.1.1, some of the synonymous groups are of exactly the same semantic, while some are of partly the same semantic. To distinguish these synonymous groups, we need to compare the semantic of each word in the synonymous group. We get the semantic representation of each word by HowNet, and then determine the synonymous relationship between each word in the synonymous group.

HowNet is a common sense knowledge base, which used the concept represented by the Chinese and English words as the object of description, and used the revealing of the relationship between concept and concept as basic content.

Take the a word in HowNet as an example, which shows that '打' and '买' are a pair of partly same synonymous group.

3) the First Series of Standardized Forms of Words with Non-Standardized Variant Form

In October 1955, the national word reform conference held in Beijing, unanimously adopted "the First Series of Standardized Forms of Words with Non-Standardized Variant Form", and recommended the immediate implementation by the press and publication department.

In December the same year, the Ministry of Culture and Committee of Cultural Reform released "the First Series of Standardized Forms of Words with Non-Standardized Variant Form". They require a national implementation from February 1956. The table included 810 groups of non-standardized variant form words. Based on the principal of simple and elementary, 810 words are selected as standardized forms of words and 1055 non-standardized ones are eliminated.

TABLE I.
EXAMPLE OF PARTLY SAME SYNONYMOUS GROUP IN SYNONYM DICTIONARY

NO.=000001
W_C=打
G_C=V
E_C=~酱油,~张票,~饭,去~瓶酒,醋~来了
W_E=buy
G_E=V
E_E=
DEF=buy|买
	NO.=015492
	W_C=打
	G_C=V
	E_C=~毛衣,~毛裤,~双毛袜子,~草鞋,~一条围巾,~麻绳,~条辫子
	W_E=knit
	G_E=V
	E_E=
	DEF=weave|辫编

After the release of the table, the number of words in Chinese characters is reduced, and the Chinese character system is also improved greatly to the direction of standardization. It also curb the font confusion phenomenon in the use of Chinese characters.

"the First Series of Standardized Forms of Words with Non-Standardized Variant Form" is undoubtedly our standard of new ones to eliminate obsolete ones. But we need to notice that it eliminate non-standardized variant form. on the basis of non-standardized variant form and traditional ones. Therefore, it can only be used as a main standard for eliminating non-standardized variant form words, but cannot be used as a normative standards to write new ones and simplified ones. Even in the field of eliminating non-standardized variant form, its role is quite limited. Since there was also some adjustments after the release of the table, some non-standardized variant form is restored in the "Simplified words table" and "Modern Chinese generic word table" released after. So if there is any inconsistency with "Simplified words table" and "Modern Chinese generic word table", the latter should be used as standard.

4) The generation of knowledge base

Regarding all the aspects above, we need to both guarantee there is enough words in knowledge base but avoid ambiguous of their meaning in the same time. The detailed steps to generate our knowledge base are as follows.

a) According to "The Dictionary of Synonym Words" made by Center for Information Retrieval of Harbin Institute of Technology.

We pick up some word groups of exactly same meaning. Then we cut the number of words of all groups to the same. Thus we get a dictionary of synonym words.

b) According to the First Series of Standardized Forms of Words with Non-Standardized Variant Form *made by China's State Language Work Committee which is published in 2001 and released in March 31,200.we get a dictionary for standardized forms of words with non-standardized variant form.*

c) Integrated the above two dictionari

.We get a rude knowledge base. We need to delete the repeated words to avoid ambiguity when encoding.

d) Segment words by ICTCLAS system, a Chinese word segment

It is made by Institute of Computing Technology, Chinese Academy of Sciences. We need to delete those that cannot be segmented correctly.

e) Delete the group that contains only one word.

f) According to "HowNet"

we can get each word's meaning and compare them with each other using the method introduced last section. We only keep those that have exactly the same meaning with the ones in the group to avoid ambiguous .

g) Using Huffman coding to encode words of each group.

After the steps above, we finally get a knowledge base of about 17000 words. The knowledge base covers almost all the synonym words and is practical to use. In this way, we can get high efficiency and more accurate when doing information hiding based on substituted conception since we have do some optimization, classification and encoding in advance.

C. An Improve Method in Encoding

Information hiding technology is mainly used in secret communication and protection of copyright for digit text. We have to come up with a method to insert as much text information as we can into a piece of article which is not very long itself. And in this same time, it is obvious that the robustness is the most important check point for information hiding technology. We can even improve the robustness at the cost of reduce capacity volume. In this way, we can fullfil the two needs by encoding the secret messages first.

Data compression is for source encoding. It makes the messages transformation quicker and higher efficient. It also increase the capacity of text information because it can decrease the loss of entropy, thus increase encoding efficiency. Its main target is to reduce the redundancy between content. Claude Elwood Shannon published the famous paper about the founded Information Theory in 1948, which is the very start of modern information theory. The Shannon's theorem, sometimes also called the noisy-channel coding theorem, establishes that for any given degree of noise contamination of a communication channel, it is possible to communicate discrete data (digital information) nearly error-free up to a computable maximum rate through the channel based in part on earlier work and ideas of Harry Nyquist and Ralph Hartley. This offers theory evidence for source encoding to compress redundancy information. In this article, we use Lempel-Ziv encoding as source encoding.

The aim of channel coding theory is to find codes which transmit quickly, contain many valid code words and can correct or at least detect many errors. While not mutually exclusive, performance in these areas is a tradeoff. So, different codes are optimal for different applications. The needed properties of this code mainly depend on the probability of errors happening during transmission. Thus the channel coding theory is developing to improve the quality of communication system. The basic method is to add redundancy symbols according to some special rule in the sending messages to ensure the reliable of the transformation. The target of channel coding theory is to construct a good encoding method to gain the minimum redundancy at the expense of the largest anti-interference performance. In this article, we use hamming code to increase the robustness of our system.

E.g. we have a piece of article in which there contains secret messages. Once the attacker knows the fact, he can do some changes to the article although he can’t get the information hide in it. He may delete some ‘的’ randomly, which will cause lose of message. Above all, it is import for us to protect the integrality at the cost of declineing the embedding rate.

1) Lempel-Ziv encoding

Kolmogoloy, the math scientist of Soviet Union proposed an encoding method by taking advantage of the construction feature of source messages. After that, two researchers from Israel, J.Ziv and A.Lempel find another method to create an even higher efficient encoding way than Huffman code. The new algorithm is totally different from Huffman algorithm and math algorithm. It also enjoys a faster compress algorithm. Now the series of algorithm is called LZ algorithm.

The detail steps are as follows.

Set the symbols of source messages as

$$A = \{a_1, a_{2,} a_3 \ldots a_k\},$$

which have totally k symbols.

Set the sequence of source messages as

$$A = \{u_1, u_{2,} u_3 \ldots u_m\},$$

which have totally m symbols.

We divide the sequence into several parts. And the purpose of the division is both ensure the minimum of the conjunction of symbols of source messages and each part is different from the other one.

We first take one symbol as the first part and keep doing segmentation. If we meet with the same symbol or symbol sequence as we meet before, we add one more symbol from the sequence next to the current one to make it different from the part already exists. We save all the parts in a table as dictionary. When the dictionary reaches a certain size, we do segmentation just according to it until the end of the sequence.

The symbol parts in dictionary are then encoded.

Set $\mathrm{M}(u)$as the number of segmentations in dictionary. If we encode the dictionary as binary code, the length of encoded word is as follow.

$$\mathrm{n} = \lceil \log M(u) \rceil$$

Lempel-Ziv code’s encoding method is very easy and fast. The same is its decoding method. It doesn’t need the whole dictionary before decode. It can set up the dictionary in the same time decode the messages. The only thing to send with messages is the size of the dictionary. Besides, the feature of Lempel-Ziv code is that the longer the source message is the higher efficiency it has. It will reach the maximum value according to Shannon's theorem

2) Hamming code

Hamming code, which is a kind of parity code, is a family of linear error-correcting codes in telecommunication, name after the inventor of it, Richard Hamming. Hamming codes can detect up to two and correct up to one bit errors. By contrast, the simple parity code cannot correct errors, and can detect only an odd number of errors. Hamming codes are special in that they are perfect codes, that is, they achieve the highest possible rate for codes with their block length and minimum distance.

In 1950, Hamming introduced the (7,4) code. It encodes 4 data bits into 7 bits by adding three parity bits. Hamming(7,4) can detect and correct single-bit errors. With the addition of an overall parity bit, it can also detect (but not correct) double-bit errors.

The code generator matrix G and the parity-check matrix H in this article are

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}_{4,7}$$

and

$$H = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}_{3,7}$$

By introducing Hamming Code to information hiding technology, we divide the binary form of secret messages into groups and then embed them into several paragraph of original messages to improve resistance to destructive. The encoding steps are as follows.

a) Convert secret messages into binary form.

b) We group every 4 binary bit and encode them by (7,4) code.

c) Embed the 7 bit into a paragraph .

The decoding steps are as follows.

d) Obtain embedded code in each paragraph.

e) Decode them by (7,4) code.

f) Correct errors if any.

III. Performance Analysis

A. *Experiments Corpus*

We need lots of text to hide secret messages in to test its security, capacity and robustness of our algorithm. Our test sources is mainly from top magazines in China, such as "Chinese Journal of Computers" and "Chinese Science". Besides, there are also some article from PRP project in SJTU.

B. *Comparison of Performance*

We will show the performance of three different methods which are all based on substituted in conception in security, capacity and robustness. Table 2 shows the statistics of capacity. The frequency of substitutions is highest for the method that based on synonym words. It is up to 36 per 1000 words. But the other two methods don't show good. It is because that the Chinese doesn't pay attention to the distinguish of "的","得" and "地".

TABLE II.

TABLE Statistics of capacity

Per words Base on	1000	2000	5000
synonym words	36	69	181
function words	15	31	80
words	4	8	11

Table 3 shows the security when hiding information. It is apparently that it is hard for the attacker to notice there are secret messages. Besides, there is no influence to the meaning of original text. As for the robustness, if the attacker knows the method how we embed secret messages, it is easy for him to delete, add or modify the functional words and auxiliary word, which will cause information lost largely. But he can only attack the text randomly if the hiding information is embedded based on knowledge. So it is more safe if embed messages according to knowledge base.

Table 4 shows the impression on system execution speed after encoding the secret messages. Although the consuming time increase after encoding, it is still accepted even it reaches a level of 50000 words.

TABLE III.

Security

	10000 words	20000 words	50000words
No encoding	19.5kb	39.1kb	97.8kb
LZ encoding	3.13kb	5.31kb	9.17kb

TABLE IV.

Impression on Speed after encoding

特征码是什么呢？比如说，“如果在第1034字节处是下面的内容：0xec，0x99,0x80,0x99,就表示是大麻病毒。”这就是特征	特征码是什么呢？比如说，“假设在第1034字节处是下面的内容：0xec，0x99，0x80,0x99,就代表是大麻病毒。”这就是特征码，一串表明病毒本身特征

码, 一串表明病毒自身特征的十六进制的字串。特征码一般都选得很长, 有时可达数十字节, 一般也会选取多个, 以保证正确判断。杀毒软件通过利用特征串, 可以非常容易得查出病毒。 （original messages）	的十六进制的字串。特征码一般都选得很长, 有时可达数十字节, 一般也会选择多个, 以确保正确判断。杀毒软件通过利用特征串, 可以非常容易得查出病毒。 （ Based on synonym words）
特征码是什么呢？比如说，“如果在第1034字节处是下面的内容 ：0xec ， 0x99, 0x80,0x99,就表示是大麻病毒。”这就是特征码, 一串表明病毒自身特征的十六进制字串。特征码一般都选得很长, 有时可达数十字节, 一般也会选取多个, 以保证正确判断。杀毒软件通过利用特征串, 可以非常容易得查出病毒 。（ Based on function words）	特征码是什么呢？比如说，“如果在第 1034 字节处是下面的内容：0xec ， 0x99, 0x80,0x99,就表示是大麻病毒。”这就是特征码, 一串表明病毒自身特征的十六进制字串。特征码一般都选地很长, 有时可达数十字节, 一般也会选取多个, 以保证正确判断。杀毒软件通过利用特征串,可以非常容易得查出病毒。 （ Based on Homophones words）

Table 5 shows the Lempel-Ziv code's impression on system capacity of hidding secret messages after encoding them. As it can be seen, it takes a much larger space to store the information without any encoding. And after introducing Lempel-Ziv code, it compress about 50% space. It gives a better performance by LZ code. Besids, the compress ration is increasing while the increase of the text length because of the LZ algorithm itself. When the text reaches 50000 words, the compression ratio is less 10%.

TABLE V.
IMPRESSION ON CAPACITY AFTER ENCODING

	10000 words	20000 words	50000 words
No encoding	0.03s	0.07s	0.14s
LZ encoding	0.06s	0.14s	0.51s
LZ & Hamming encoding	0.15s	0.3s	0.81s

CONCLUSION

The analysis of information hiding technology is significant to security insurances of communication and copyright protection of published digit text. It also has a broad application prospect. However, most studies nowadays focus on hiding technology based on the format of a document. In this article we analyze three kinds of text information hiding methods based on natural language, and propose an efficient information hiding algorithm for text based on substituted conception. Experiment shows it has a good performance. Moreover, in this article we also discuss attack resist method in text hiding.

ACKNOWLEDGMENT

This paper is supported by the National Key Basic Research Program of China (2013CB329603), the National Natural Science Foundation of China (61272441,61171173) and the Opening Project of Key Lab of Information Network Security of Ministry of Public Security (The Third Research Institute of Ministry of Public Security), whose number is C12609.

REFERENCES

[1] J.T.Brassil,S.Low, N.F.Maxemchuk.Copyright Protection Electronic Distribution of Text Documents.Proceedings of the IEEE,1 999,87(7)

[2] Katzenbeisser'S.C..Principles of Steganography.In:Tcchniques for Steganographyand Digital Watermarking,Boston,2000,17-41

[3] M.Atallah,C.McDonough,S.Nirenburg,et a1.Natural Language Processing for Information Assurance and Security: An Overview and Implementations.In: Proceedings 9th ACM／SIGSAC New Security Paradigms Workshop,Ireland 2000,51-65,

[4] Mikhail,J.Atallah et al,Natural Language Watermarking:Design,Analysis,and a Proof of Concept Implementation. In:Information Hiding 200l,2001,185.199

[5] M.S.Kankanhalli, K.F. Hau. Watermarking of Electronic Text Documents.In: Electronic Commerce Research. Netherlands:Kluwer Academic Publishers, 2002,169-1 87

[6] T.Mark,I.George,Marc Rennhard.A Practical and Effective Approach to Large-scale Automated Linguistic Steganography.In:Information Security: Fourth International Conference.Heidelberg:Springer Berlin,2004,156-165

[7] Mei Jiaju, Zhu Yiming, "The Dictionary of Synonym Words",Shanghai Lexicographic Publishing House,1983

[8] stefan kazenbeisser,Fabien Petitolasa ,Information Hiding Techniques for Steganography and Digital Watermaking,EDPACS,Volume 28, Issue 6, 2000

[9] Beijing Language Institute, Chinese vocabulary statictics and analysis,Foreign Language Teaching and Research House,1985.4

[10] [Dong Zhendong,HowNet.http://www.keenage.com/ 2006.9.28

[11] the First Series of Standardized Forms of Words with Non-Standardized Variant Form, the Ministry of Culture and Committee of Cultural Reform,2002,3,31

[12] Abdelrahman Desoky, Noiseless Steganography: The Key to Covert Communications,CRC press Feb,2012

[13] A. Desoky, Nostega: A Novel Noiseless Steganography Paradigm, Journal of Digital Forensic Practice,Vol.2, 132-139 March,2008.

[14] ZHOU, X., WANG, S., XIONG, S., YU, J.. Attack Model and Performance Evaluation of Text Digital Watermarking. Journal of Computers, North America, 5, dec. 2010.

[15] ZHANG, X., ZHAO, G., NIU, P.. A Novel Approach of Secret Hiding in Webpage by Bit Grouping Technology. Journal of Software, North America, 7, nov. 2012.

[16] DAI, W., YU, Y., DAI, Y., DENG, B.. Text Steganography System Using Markov Chain Source Model

and DES Algorithm. Journal of Software, North America, 5, jul. 2010.

Gongshen Liu. Shandong, China. Feb. 12th, 1974. He got his Ph.D. on computer science from Shanghai Jiao Tong University (SJTU), 2003; M.A. on computer science from Shandong University, 2000 and B.A. on computer science from Shandong University of Technology 1997.

He is an Associate Professor of School of Information Security Engineering of SJTU. He has many research experiences in the field of Natural Language Processing, Social Network and Content-based Security, some of which are published in International conferences and journals, such as China Communication, Journal of Systems Engineering and Electronics and so on.

Dr. Liu is the member of ACM, China Computer Federation and Chinese Information Processing Society of China.

Xiaoyun Ding. Shanghai, China, 1988.11, Graduate of School of Information Security Engineering of Shanghai Jiaotong University. Research in web content security and text classification and steganography.

She is now major in content security lab of SJTU in Shanghai as a Graduate student. She has just taken part in ISISE 2012 conference for a paper about information hiding technology.