An Information Hiding method for Text by Substituted Conception

Su Bo, School of Information Security Engineering, Shanghai Jiao Tong University Shanghai China
Ding Xiaoyun, School of Information Security Engineering, Shanghai Jiao Tong University Shanghai China
Liu Gongshen , School of Information Security Engineering, Shanghai Jiao Tong University Shanghai China
Zhang Hao, Xie Jing School of Information Security Engineering, Shanghai Jiao Tong University

***Abstract*—After analysis of various existing text-based steganography techniques, an efficient information hiding algorithm for text based on substituted conception is proposed in this paper. The algorithm achieves information hiding by replacing the substituted units in original text with similar conception. It takes advantage of a user-defined knowledge base, according to which the secret message can be encoded and then embedded into stego-text. It's proved by the experiments that the algorithm proposed in this paper performs well in two aspects, security and capacity. Last but not the least, in order to enhance the robustness of the algorithm; we also proposed an improved encoding algorithm to prevent information loss caused by attack.**

Keywords: substituted conception; information hiding; text; encoding

I. INTRODUCTION

Information hiding technology means the concealing of the information itself and its location. The largest and deepest research is based on the image as carrier to hide open information and digital watermark. This is mainly due to that image has large redundant information capacity, and in the other aspect image process is more intuitive. But text does not contain enough redundant information. Information hidden in text would be very difficult.

The first academic information-hiding in the text work that is published in last century, which is in the era of computer, called "Mimic functions" by Peter Wayner in 1992. Then, in 1997 Chapman and Davida have introduced a steganographic scheme consisting of two functions called NICETEXT and SCRAMBLE that uses a large dictionary. The approach is also known as NICETEXT. Mikhail J. and M. Attalla of Purdue University then in 2000 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 changes the attributes of the original text to embedding secret information in while keep the meaning of original text. Finally, numerous approaches were introduced and published after 2005[12][13].

It is generally considered that there are three methods to realize text hiding algorithm.

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 plan text, the format information in the formatted text contains more redundancy. The second one is based on the syntax of text. It achieves the target by changing the syntactic structure of sentences. It is commonly used in mapping mode, move the position of adjunct join the form of the subject, changing the active and passive form, changing the statement sequence. The third one is based on the semantic of the text. Many researchers are working on it now, especially on doing replacement based on synonyms. It performs well on the consistency of the text and encoding of the synonyms of the text. Reference [2] works deeply on the synonym replacement method. However, the research about the method based on the synonym is far from enough in Chinese language.

In this article, we proposed an information hiding method based on the substituted conception. In the second chapter, we combine the encoding of the secret information and looking up of a self-definition knowledge base to realize text hiding. We will introduce the definition of the substituted conception, and bring in knowledge base and coding algorithm next. We integrated the above skills into one system platform. In the second half of the second chapter we will go deeply describing how to use this conception and the detailed steps. At the end of this chapter we will present an improved coding algorithm——hamming code to make the system stronger. Performance testing of the system platform will be listed in the third chapter, for readers to analyze and compare.

II. THE INFORMATION HIDING METHOD BASED ON THE SUBSTITUTED CONCEPTION

Substituted conception is defined as: for some certain parts in the text, we can find some units from 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 give rise to ambiguity or error.

The synonym is a kind of substituted unit. There are a lot of researches of the substituted method based on synonyms in the foreign country. However, for the Chinese characteristics, 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

978-0-7695-4951-4/13 $26.00 © 2013 IEEE
DOI 10.1109/ISISE.2012.36

unit definition, Knowledge base generation and encoding algorithm of information hiding system platform.

The flow chart of our platform is as Figure1 shows.

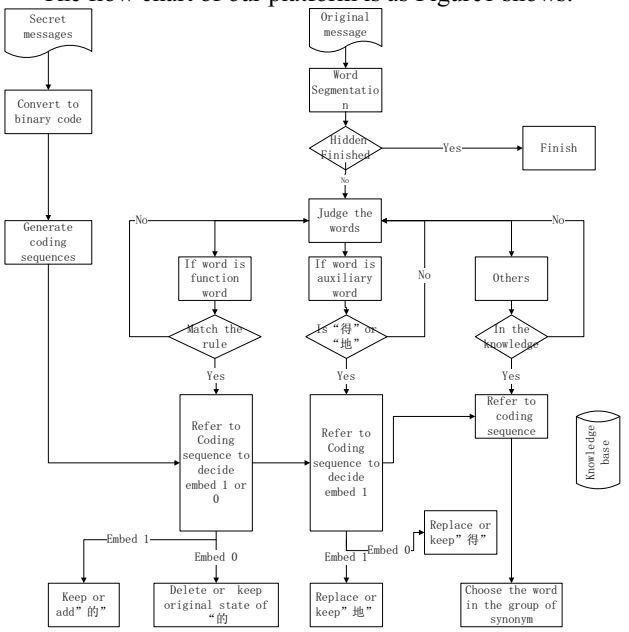

Figure 1. Figure 1. The diagram of information hiding by substituted conception

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 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 information in 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 the second half of the chapter, 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 knowledge base. If it is in the knowledge base, according to the order number and location of the group of words in the knowledge base and the embedded algorithm coding and to obtain the result of binary string. A Repeat these steps until the end of the flag.

This article uses the "Hownet" and "The Dictionary of Synonym Words" as mentioned dictionary. Please refer to in-depth study on this issue in chapter 2.2

2) Use Function Words as Substituted Conception

From the language statistical study, we find that function words such as '的', '了', '是' enjoy a high frequency in Chinese. When the original text 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. 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 concluded several rules that can be embedded with hidden information. In these rules, the '的' can be added or deleted without rising other people's awareness.

a) rule 1 '的' can be removed when using monosyllabic adjective as attribute

e.g.：新发现—新的发现

b) rule 2 nouns as attribute

e.g.：玻璃门—玻璃的门

c) rule 3 pronouns as attribute

e.g.：我祖国—我的祖国

d) 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.

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

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.

a) Standardized forms of words and non-standardized ones. In its development process, there appears a large amount of standardized forms of words and non-standardized ones." 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.

b) 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 while ‘得’ follows the verb or adjective when it is used as structural particle. 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". It gets 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 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.

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.A, 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 dictionary, 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 in table 1.

TABLE I. EXAMPLE OF PARTLY SAME SYNONYMOUS GROUP IN

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

SYNONYM DICTIONARY

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".

It 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.W*e 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 dictionaries.*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 which 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. And in this way, 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.

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. The Hamming (7,4) 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.

1) Convert secret messages into binary form.
2) We group every 4 binary bit and encode them by (7,4) code.
3) Embed the 7 bit into a paragraph .
4) The decoding steps are as follows.
5) Obtain embedded code in each paragraph.
6) Decode them by (7,4) code.
7) Correct errors if any.

III. PERFORMANCE ANALYSIS

A. Experiments corpus

We need lots of text to hide secret messages in to exam its security, capacity and robustness of our algorithm. Our sources for test 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". The result shows that there is a certain captivity to conceal information.

TABLE II. 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. 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 III. SECURITY

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的字串。 （original messages）	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的十六进制的字串。 （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。 （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。 （Based on Homophones words）

IV. CONCLUSION

The analysis of information hiding technology is very important in ensure the security of communication and the protection of the publish of digit text. It also has a broad application prospect. The study now focus on hiding technology based on the format of a document but there is little based on natural language. In this article we go deep into three methods which are all based on substituted in conception and propose the way to generate knowledge base. Experiment shows a good performance. What's more, this article also discusses the method of attack resistanceAcknowledgment

ACKNOWLEDGMENT

This paper is supported by the National Engineering Lab of Information Content Analysis Technology (GT036001), the National Natural Science Foundation of China (61272441，61171173).

REFERENCES

[1] J.T.Brassil，S．Low, N.F.Maxemchuk．Copyright Protection Electronic Distribution of Text Documents．Proceedings of the IEEE，1999，87(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. Watemarking 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