Natural Language Information Hiding Based on Chinese Mathematical Expression

Yuling Liu[1], Xingming Sun[1], Ingemar J. Cox[2], and Hong Wang[1]
(Corresponding author: Xingming Sun)

School of Computer and Communication, Hunan University[1]
Changsha, 410082, Office 508, P.R.China (Email: yuling_liu@126.com)
Department of Computer Science and Department of Electronic and Electrical Engineering[2]
University College London

(Received Mar. 4, 2007; revised and accepted Mar. 7, 2008)

Abstract

A novel method of natural language information hiding is presented based on Chinese mathematical expression. The method can convert Chinese text into bit string by using the thought of Chinese mathematical expression, and then embed the secret message into the text by utilizing the replacement of synonyms, variant forms of the same word, the adding or deleting of the empty words, and the shift conversion of the sentence. Experimental results show that the method can achieve a faster conversion of the text into the binary string, a better degree of information-carrying capacity and a better result with the imperceptibility.

Keywords: Chinese mathematical expression, natural language information hiding, shift conversion, the empty word, variant forms of the same word

1 Introduction

With the development of Internet and P2P networks, people can publish their works, transmit important messages, and do trade network conveniently. Digital text is the most common and frequently used carrier (digital data) and has diverse forms, such as Webpage, e-mail, kinds of formatted text files including PS, PDF, DOC, TXT, and so on. Some data shows that the digital text's transmission accounted for a very large proportion of network traffic, so it is much easier to transmit the secret message in text.

Text information hiding is to use text as the medium to hide information. The early methods of text information hiding are based on the physical formatting of text. Due to those methods exploited tolerances in typesetting by making minute changes in line placement and kerning, making them vulnerable to simple reformatting and OCR attacks [15], some researchers introduced some natural language processing techniques into text information hiding, which is natural language information hiding. However, in natural language information hiding methods, to encapsulate the hidden information depends only on bit string and not on the text. In order to convert the text to bit string, much work has been done. The first method is derived from the sentence's corresponding tree structure [2, 3]. To each sentence s_i, there exists a corresponding tree T_i that represents either s_i syntactically or semantically. The nodes of the tree T_i are labeled in pre-order traversal. Then, a node label j is converted to 1 if $j+H(p)$ is a quadratic residue modulo p, and to 0 otherwise, where p is a secret key and $H(\cdot)$ is a one-way hash function. A node label sequence B_i is then generated according to a post-order traversal of T_i. The second method is the binary tree encoding methodology of the synonym-based algorithm [5]. All words in all synonyms are translated to bit strings, and then divide to two sub-groups by bit operation and quadratic residue table lookup. Each sub-group should be divided into two sub-groups until the sub-group has one synonym or no bit operations can be used. The third method is based on the sentence length [6]. The number of words in a sentence can be represented in the binary form.

Naturally, such methods can convert the text into the binary strings successfully, but this paper presents a novel encoding method based on the mathematical expression of a Chinese character, which can convert each Chinese character into a binary bit according to the structure relation of components of the Chinese character. Then we introduce the replacement of synonyms, variant forms of the same word, and the adding or deleting of empty words. Additionally, the shift conversion of the sentence also is developed.

The rest of the paper is organized as follows. Section 2 describes the related works, such as natural language information hiding and thought of the mathematical expression of a Chinese character. Section 3 provides a detailed description of our proposed scheme, and then the experimental results are reported in Section 4. Conclusions and future directions are given in Section 5.

2 Related Works

2.1 Natural Language Information Hiding

Information hiding of natural language is one of text information hiding based on natural language processing techniques, which is becoming a hot spot. Publicly available methods for natural language information hiding can be classified into two groups. The first group of methods is based on generating a new text document for a given message with or without the given cover document [4, 9]. The other group of methods is based on modifying a given cover document to embed the message in it. Recent work has fallen into the second type, and the modifying involves the word level, the sentence level and the entire paragraph. Synonym replacement method is a more mature technique at the word level [1, 5, 14, 15]. Syntactic transformation based method has been proposed at the sentence level [2, 6, 8, 13]. Furthermore, M. J. Atallah et.al in CERIAS has presented a natural language information hiding technique based on TMR [3] for the entire paragraph.

2.2 The Mathematical Expression of a Chinese Character

Based on the knowledge of the structure of Chinese characters, each Chinese character can be denoted by a mathematical expression [11]. The operands are components of Chinese characters and the operators are 6 spatial relations of components. Some definitions are given below.

Definition 1. *A basic component is composed of several strokes, and it may be a Chinese character or a part of a Chinese character.*

Definition 2. *A compound component is composed of two or more than two basic components.*

Definition 3. *An operator is the location relation between the components. Let A,B be two components, A* lr *B, A* ud *B, A* ld *B, A* lu *B, A* ru *B and A* we *B represent that A and B have the spatial relation of left-right, up-down, left-down, left-upper, right-upper, and whole-enclosed respectively. An intuitive explanation of six operators sees Figure 1.*

Definition 4. *The priority of the six operators defined in Definition 3 is as following: ( ) is the highest (1);* we, lu, ld, ru *are in the middle (2);* lr, ud *are the lowest (3); the operating direction is from left to right.*

Using the selected 600 basic components and the six operators defined above, we can express all the 20902 CJK Chinese characters in UNICODE 3.0 in their mathematical expressions. The theory has been applied successfully in fonts automation, Chinese information transmission, discovery of the structure of Chinese characters. It can also be applied extensively to many areas such as typesetting, advertising, packing design, virtual library, network transmission, pattern recognition and Chinese mobile communication. Furthermore, it has been applied to text information hiding [12].

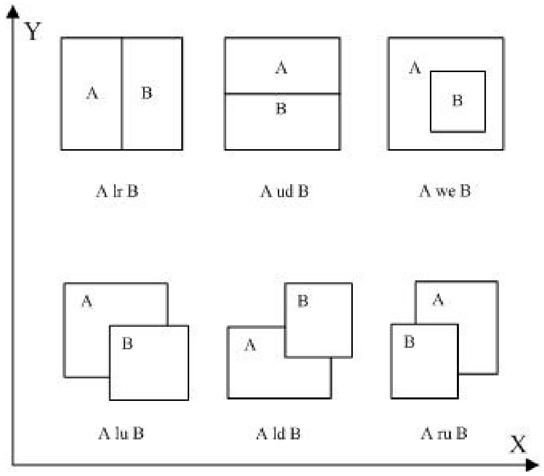

Figure 1: Intuitive explanation of the defined operators

3 Proposed Schemes

3.1 Text Binarization

Since the hidden information cannot be embedded directly in the natural language text, it is necessary to convert the text into the binary string before the embedding process. For Chinese, this paper presents a novel text binarization method based on the thought of the mathematical expression of a Chinese character.

Let c be a Chinese character, and Γ be a set of the 20902 CJK Chinese characters in UNICODE 3.0. If $c \in \Gamma$, $math(c)$ represents the mathematical expression of the Chinese characters c. Furthermore, $math(c)$ can be represented in a binary tree, in which the basic components are the leaves, and the operators are the intermediate nodes. When c is a basic component, c cannot be represented in binary tree. When c is a compound component, $math(c)$ can be represented in the form of $lopr$, where op is an operator of the root in a binary tree, l and r are the left sub-tree and the right sub-tree of the operator op in the mathematical expression respectively. For example, $c =$ 在, $math(c) = 498lu28$, is shown in Figure 2.

For each Chinese character c, the conversion process into one bit involves four steps:

1) Represent c mathematically and get the expression *math(c)*.

2) Extract the operators from the expression by deleting all the operands, all the brackets and their contents.

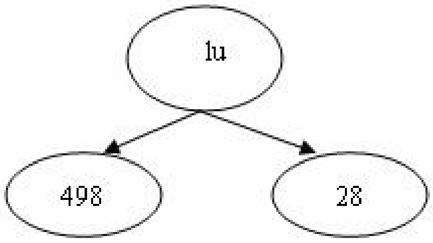

Figure 2: The binary tree of Chinese character “在”

3) Compare the priority of the operators according to Definition 4, and get the operator *op* with the lowest priority, which will be the root of the binary tree. All the operands and operators to the left of *op* will be the left sub-tree, and those to the right of *op* will be the right sub-tree.

4) Compute the sum of the operands in the left sub-tree (*sum(l)*) and the sum of the operands in the right sub-tree (*sum(r)*) respectively, and derive the bit value of c(*mark(c)*) from the following equation:

$$mark(c) = \begin{cases} 0, & c \in \Gamma\ AND\ sum(l) < sum(r) \\ & OR\ c\ isn't\ in\ \Gamma \\ & \\ 1, & c \in \Gamma\ AND\ suml \geq sum(r) \end{cases}$$

3.2 Some Useful Linguistic Transformations

Unlike the image, video and audio carrier, text has little redundant information. However, there are some redundancies by using linguistic transformation in natural language text. The proposed linguistic transformations involve the word level, the sentence level and the entire chapter. After the conversing the text into the binary strings, we exploit some useful transformations, such as synonym replacement, the replacement of variant forms of the same word, the adding or deleting of the empty words, and the shift conversion of the sentence.

Definition 5. *Broad synonyms: Absolute synonyms are the words which have the same meaning of all the senses in the same language. However, the number of absolute synonyms in Chinese is relatively small to nonexistent, so we define the broad synonyms, which exist in the dictionaries and have the same meaning in one or some senses.*

Definition 6. *Variant forms of the same word are the different writing words, which are homophonic (sound, rhyme, tone), and have the same meaning.*

Definition 7. *The empty words are the words, which have empty meaning but use frequently, such as the auxiliary word “的 ”、“地 ”、“得 ” and so on.*

Definition 8. *Shift conversion of the sentence is referring to that one part of the sentence, such as a word or a phrase can be placed in the different location in the sentence.*

Descriptions for the aforementioned transformtions are given below.

1) The replacement of synonyms and variant forms of the same word. The synonyms and variant forms of the same word have the same meaning, so we can achieve the goal of embedding the secret message into the text by the replacement of them. We design a thesaurus based on words from the existing word lists of synonyms and variant forms of the same word, and moderately refine and augment or reduce manually by applying a Chinese lexical analysis system called ICTCLAS [7] to the words. For example, “包括” and “包含” is a synonym cluster; “賢惠” and “賢慧” is a cluster of variant forms of the same word.

2) The adding or deleting of the empty words. Due to the empty meaning, the empty words can be selected to embed the secret message through adding or deleting them without changing the meaning of the text. In this paper, we just select the deleting the DE word in DE phrase. For example, the sentence “主席向世界各國的朋友們問好.” can be modified to “主席向世界各國朋友們問好.” after deleting the DE word.

3) The shift conversion of the sentence. Based on the characteristics of Chinese grammar, we rewrite the sentence by exploiting the shift conversion rules of the sentences that some linguists have summarized. For example, the preposition phrase can be placed before or after the verbs, even before the noun of agent. We just select the preposition “在”、“對”、“爲”、“從”、“向”、“用”.

3.3 Embedding Algorithm

Let S be the sentences that have been selected for embedding the information in the text. The embedding process is as follows.

1) Convert each sentence s_i in S into a bit string B_i according to Section 3.1, and then repeat the following Steps 2 - 5 until all the secret message has been embedded.

2) Text preprocessing: lexical analysis and part of speech tagging.

3) We pick the information-carrying β bits within B_i according to the key p, and verify if β bits match the desired values. If so return, else go to next step.

4) For each transformation in aforementioned transformations;

Table 1: The applicable capacity results of three transformations

The Methods	Applicable term	The percent
The replacement of synonyms and variant forms of the same word based	26715	9.10%
The deleting of the empty words based	17059	5.81%
The shift conversion rules based	6668	2.27%
The total	50442	17.18%

a. If there exists any synonym or variant forms of the same word in the sentence s_i, replace it to get the sentence s'_i, and convert s'_i into B'_i. Then verify if the relevant bits of B'_i match the desired values. If so go to next step, else go to (b).

b. If there exists any DE phrase in the sentence s_i, delete DE to get the sentence s'_i, and convert s'_i into B'_i. Then verify if the relevant bits of B'_i match the desired values. If so go to next step, else go to (c).

c. If there exists any preposition phrase that satisfies the shift conversion rules in the sentence s_i, re-write s_i to get the sentence s'_i, and convert s'_i into B'_i. Then verify if the relevant bits of B'_i match the desired values. If so go to next step, else skip s_i.

5) Obtain the transformed sentence s'_i, and select the next sentence s_{i+1}.

3.4 Extracting Process

The extracting process is similar to the embedding process except that reading the bit value rather than modifying the sentences. We just describe it briefly as follows.

Let S' be the sentences that have been selected for embedding the information in the text. The extracting process repeats the following Steps 1 - 2 until the last sentence in S'.

1) Convert each sentence s_i in S' into a bit string B_i according to Section 3.1.

2) Pick the information-carrying β bits within B_i according to the key p, and estimate whether the corresponding terms to β bits satisfy the aforementioned transformation or not. If so, read the β bits, else skip s_i.

3) Link the β bits and obtain the secret message.

4 Experiments

We have implemented the binarization of all the 20902 CJK Chinese characters in UNICODE 3.0, thus we can lookup the corresponding bit value directly while embedding and extracting information. Compared with other methods based on morphological, syntactical, or semantical parsers, our method can convert text into bit strings more quickly.

In order to demonstrate the possibility of applying aforementioned transformations on a real-life test case, we select PFR corpus as the training data [10]. The PFR corpus is from the corpus of People's Daily in 1998, and has tagged the part of speech. The free part of PFR corpus is the January publication in 1998. We just pick the first ten publication days of the free part, and the count of word is 293670. In order to embed the secret message, the synonyms and variant forms of the same word, DE word, and the aforementioned prepositions are the candidate information-carrying terms, which are applicable terms. The applicable capacity results of three transformations list the following Table 1. It was observed that the single method had less capacity as compared to the combination with several methods.

It is hard to draw any truly data of capacity from the previous methods, because several reasons will be considered, such as the corpus used, the native difference between English and Chinese, and so on. Below is a summary of the previous work of natural language information hiding in Table 2.

Additionally, the essential of information hiding is to keep the cover documents imperceptible. Natural language information hiding should embed the secret message without changing the meaning of the text. Similar to the PSNR used to measure the perceptual transparency in the image or video information hiding, we introduce the imperceptibility degree (for short IMD) in the natural language information hiding. $IMD = \frac{\sum A_i}{\sum T_i}$, where A_i is the total number of successfully transformed terms for each sentence. And T_i is the total number of applicable terms for each sentence. Combining automatic evaluation with human interaction, we can measure whether the terms are successfully transformed or not. In the method of the replacement of synonyms and variant forms of the same word, Chinese automatic disambiguating based on Chinese lexical analysis is introduced. However, because of the limit of the current Chinese information processing techniques, the methods based on deleting the empty words and shift conversion introduce human labor. Table 3 lists the experimental result of transparency. Since the complicate Chinese automatic disambiguating techniques are not considered in the first method, the IMD of the

Table 2: Summary of the previous work of natural language information hiding

Scheme	The descriptions
[1, 2]	Generating a new text document
[3, 8, 13, 15]	Replacement of synonyms
[4, 6, 10, 14]	Transformation of sentence structures
[7]	Transformation of TMR
Our method	Integrating the methods based on the replacement of synonyms and variant forms of the same word, deleting the empty words, and shift conversion

Table 3: The result of imperceptibility

The Methods	Applicable terms	Successfully transformed terms	IMD
The replacement of synonyms and variant forms of the same word based	26715	23106	0.86
The deleting of the empty words based	17059	9710	0.57
The shift conversion based	6668	2497	0.37

first method in the first row in Table 3 is much higher than the latter two ones.

5 Conclusions and Future Works

This paper proposed a text binarization method, which is based on Chinese mathematical expression. The method can achieve the goal of converting Chinese text into bit string conveniently, and the conversion is one-way and secure. Moreover, some useful linguistic transformations suitable for Chinese textual messages were introduced to embed the secret information. These available transformations is context based rather than format based, furthermore, they can not only enlarge the capacity, but also enhance the imperceptibility of the text information hiding.

Acknowledgements

This paper is supported by Key Program of National Natural Science Foundation of China (No.60736016), National Basic Research Program of China (No.2006CB303000), and National Natural Science Foundation of China (No.60573045).

References

[1] M. J. Atallah, C. J. McDonough, V. Raskin, and S. Nirenburg, “Natural language processing for information assurance and security: An overview and implement- Ations,” *The Workshop on New Paradigms in Information Security (NSPW 2000)*, pp. 51-56, Ireland, Sep. 2000.

[2] M. J. Atallah, V. Raskin, M. Crogan, C. Hempelmann, F. Kerschbaum, D. Mohamed, and S. Naik, “Natural language watermarking: Design, analysis, and proof-of-concept implementation,” *The Fourth International Information Hiding Workshop (IH 2001)*, LNCS 2137, pp.185-199, 2001.

[3] M. J. Atallah, V. Raskin, C. F. Hempelmann, M. Karahan, R. Sion, U.Topkara, and K. E. Triezenberg, “Natural language watermarking and tamperproofing,” *The Fifth International Information Hiding Workshop (IHW 2002)*, LNCS 2578, pp. 196-212, 2002.

[4] M. Chapman, G. Davida, and M. Rennhard, “A practical and effective approach to large scale automated linguistic stegano graphy,” *Information Security Conference*, pp. 156-165 Malaga, Spain, 2001.

[5] Y. L. Chiang, L. P. Chang, W. T. Hsieh, and W. C. Chen, “Natural language watermarking using semantic substitution for Chinese text,” *International Workshop on Digital Watermarking (IWDW 2003)*, LNCS 2939, pp. 129-140, 2003.

[6] G. Gupta, J. Pieprzyk, and H. X. Wang, “An attack-localizing watermarking scheme for natural language documents,” *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, pp. 157-165, Taipei, Taiwan, 2006.

[7] ICTCLAS, 2006-09. (http://mtgroup.ict.ac.cn/zhp/ICTCLAS.htm)

[8] B. Murphy, *Syntactic Information Hiding in Plain Text*, Master’s thesis, CLCS, Trinity College Dublin, 2001.

[9] W. Peter, “Mimic functions,” *Cryptologia*, vol. XVI, no. 3, pp. 192-213, 1992.

[10] PFRCorpus, 2006-11. (http://www.icl.pku.edu.cn)

[11] X. M. Sun, H. W. Chen, L. H. Yang, and Y. Y. Tang, "Mathematical representation of a Chinese character and its applications," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 16, no. 8, pp. 735-747, 2002.

[12] X. M. Sun, G. Luo, and H. J. Huang, "Component-based digital watermarking of Chinese texts," *The 3rd International Conference on Information Security*, pp. 76-81, 2004.

[13] M. Topkara, U. Topkara, and M. J. Atallah, "Words are not enough: Sentence level natural language watermarking," *The ACM Workshop on Content Protection and Security (in conjunction with ACM Multimedia)*, pp. 37-46, Santa Barbara, CA, 2006.

[14] U. Topkara, M. Topkara, and M. J. Atallah, "The hiding virtues of ambiguity: Quantifiably resilient watermarking of natural language text through synonym substitutions," *The ACM Multimedia and Security Workshop*, pp. 164-174, Geneva, Switzerland, Sep. 2006.

[15] K. Winstein, *Lexical Steganograph- y Through Adaptive Modulation of The Word Choice Hash*, 1998. (http://www.imsa.edu/keithw/tlex/)

Yuling Liu received the B.S. degree in Computer Science and Technology from College of Computer and Communication in Hunan University, Hunan Province, China. Then, she started the successive postgraduate and doctoral programs of study from 2003. Currently, she has been PH.D degree candidate in computer application of Hunan University. Her research interests include information security, text steganography and natural language processing.

Xingming Sun received the B.S. degree in mathematics from Hunan Normal University, China, in 1984, the M.S. degree in computing science from Dalian University of Science and Technology, China, in 1988, and the Ph. D. degree in computer science from Fudan University, China, in 2001. He was a visiting professor in University College London, UK in 2007. He is currently a professor in School of Computer and Communication at Hunan University, China. He is also a visiting professor in the Department of Computer Science at the University of Warwick, U.K. His research interests include network and information security, digital watermarking, database security, and natural language processing.

Ingemar J. Cox received his B.S. degree from University College London and Ph.D. degree from Oxford University. He has worked for AT& T Bell Labs and NEC Research Institute and is currently Professor and Chair of Telecommunications in the Departments of Electronic Engineering and Computer Science at University College London. He has worked on problems to do with stereo and motion correspondence and multimedia issues of image database retrieval and watermarking. From 1997 till 1999, he served as Chief Technical Officer of Signafy Inc., a subsidiary of NEC responsible for the commercialization of watermarking.

Hong Wang received his bachelor's degree from Central South University, China in 2000. He started to work there in the same year. Currently, he is a Master student in computer science at Hunan University, China. His areas of research interest include information hiding, network security and natural language processing as well as the practical implementation and evaluation of a range of related techniques.