Natural Language Watermarking Using Semantic Substitution for Chinese Text

Yuei-Lin Chiang, Lu-Ping Chang, Wen-Tai Hsieh, and Wen-Chih Chen

Advanced e-Commerce Technology Lab., Institute for Information Industry, 17FL.-A, No.333, Sec.2, Duenhua S. Rd., Taipei, Taiwan, 106, R.O.C.
{ylchiang, clp, wentai, wjchen}@ iii.org.tw

Abstract. Numerous schemes have been designed for watermarking multimedia contents. Many of these schemes are vulnerable to watermark erasing attacks. Naturally, such methods are ineffective on text unless the text is represented as a bitmap image, but in that case, the watermark can be erased easily by using Optical Character Recognition (OCR) to change the representation of the text from a bitmap to ASCII or EBCDIC. This study attempts to develop a method for embedding watermark in the text that is as successful as the frequency-domain methods have been for image and audio. The novel method embeds the watermark in original text, creating ciphertext, which preserves the meaning of the original text via various semantic replacements.

1 Introduction

The Internet is a two-edged sword: While the Internet offers authors the chance to publish their works worldwide, it simultaneously creates the risks of unauthorized publishing and copying. Copyright is not lost when publishing via the Internet. However, the right alone generally is insufficient to protect intellectual property. The Internet is becoming more important than the traditional medium of CD-ROM. However, effective protection is required owing to inexpensive CD writers and the mass-production of pirated CDs overseas.

The principle of digital watermarking is widely applied in bank notes. Insignificant characteristics are sufficient to verify originality and identify counterfeits. Modern digital watermarks are a widespread method for making a published work unique and thus enabling its automatic identification. Digital watermarks enable author rights to be verified in the event of unauthorized publication or copying. Several methods exist for digitally watermarking images and music, all of which involve imprinting video- and audio-files with an innocuous mark. This mark of authorship simplifies identification, and enables pirates to be caught and sentenced easily and inexpensively.

Many techniques have been proposed for watermarking multimedia contents. However, many of these techniques are defective in that they are vulnerable to watermark erasing attacks. The most successful of these techniques operate in the frequency domain [1], [2], [6]. Naturally, such methods do not work on text unless the text is represented as a bitmap image (with, for instance manipulation of kerning and/or spacing to hide the watermark), but in that case the watermark easily can be erased by

T. Kalker et al. (Eds.): IWDW 2003, LNCS 2939, pp. 129–140, 2004.

using OCR (Optical Character Recognition) to change the representation of the text from a bitmap to ASCII or EBCDIC. This study attempts to develop a method for embedding watermarks in the text that is as effective as the frequency-domain methods have been for image and audio. The proposed method embeds watermark in original text, creating a ciphertext, which preserves the meaning of the original text through various synonym replacements. The proposed is context based rather than format based. The proposed method thus can enhance the protection of text contexts.

The rest of the paper is organized as follows. Section 2 describes related works, then Section 3 provides a detailed description of our method for Chinese NLP Watermarking. Experimental results then are reported in Section 4, and future directions are given in Section 5.

2 Related Work

2.1 Natural Language Watermarking

Mikhail Atallah and Victor Raskin proposed a technique for information hiding in natural language text. Moreover,[7], [8] established the basic technique for embedding a resilient watermark in NL text by combining a number of information assurance and security techniques with the advanced methods and resources of natural language processing(NLP). A semantically based scheme significantly improves the information-hiding capacity of English text by modifying the granularity of meaning of individual terms/sentences. However, this scheme also suffered the limitations：

1. The NLP technique is suitable for English. However, for Chinese the NLP technique is differs from the English domain. The technique thus requires adapting to Chinese text contexts.
2. The technique was merely conceptual. The details of the technique were not clarified. Such details include how to select the candidate terms/sentences for embedding watermark bits in terms/sentences, the encoding/decoding algorithm, the synonymous change algorithm, and so on.
3. The technique was applicable to long text because it ensured the embedding of just one bit of the watermark bit string in each terms/sentences and required a marker sentence for each watermark-bearing sentence, thus effectively reducing the bandwidth to 0.5.

2.2 Quadratic Residue

This study uses a theorem - Euler's Quadratic Residue Theorem, to help embed watermark in the text. This theorem is important in determining whether the integer x is the square of an integer modulo p. If there is an integer x exists such that $x^2 \equiv q (\mathrm{mod} p)$, then q is said to be a quadratic residue (mod p). If not, q is said to be a quadratic non-residue (mod p). Hardy and Wright [4] use the shorthand notations $q \mathrm{R} p$ and $q \mathrm{N} p$, to indicate whether q is a quadratic or non-quadratic residue, respectively. For exam-

ple, $4^2 \equiv 6 (\mathrm{mod}\, 10)$, so six is a quadratic residue (mod 10). The entire set of quadratic residues (mod 10) is given by 1, 4, 5, 6, and 9, making the numbers 2, 3, 7, and 8 the quadratic non-residues (mod 10). Figure 1 shows the entire set of quadratic residues (mod 10).

$$1^2 \equiv 1 \pmod{10} \quad 2^2 \equiv 4 \pmod{10} \quad 3^2 \equiv 9 \pmod{10}$$
$$4^2 \equiv 6 \pmod{10} \quad 5^2 \equiv 5 \pmod{10} \quad 6^2 \equiv 6 \pmod{10}$$
$$7^2 \equiv 9 \pmod{10} \quad 8^2 \equiv 4 \pmod{10} \quad 9^2 \equiv 1 \pmod{10}$$

Fig. 1. Quadratic residues (mod 10)

3 System Architecture

Owing to the imperfect nature of human hearing and sight, watermarks can be used to hide information in texts. For example, when a green pixel is placed in the middle of a group of red pixels in a graph, human sight will not recognize the green pixel. Similarly, information can be hidden in music files using psychoacoustics. These flaws in OMIT human sensory organs cause redundancies in information hiding. These redundancies can be used to recognize the legality of the file modification, or whether the encrypted message is already hidden. However, this redundancy is smaller for natural language texts than for other texts. For example, the meaning may differ when some terms are modified or when term sequences in the text are changed. Therefore, this study proposes an approach for enabling the natural language based watermark. This approach can be used to embed a watermark, such as copyrights, in the text. The meaning of the text remains unchanged after encrypting, while the property right is protected.

This study uses the NLP watermarking definition described in [8]. The definition is：

NLP for Watermarking. Let T denote a natural language text, and let W be a string that is much shorter than T. This study attempts to generate natural language text T' such that:

1. T' has essentially the same meaning as T;
2. T' contains W as a secret watermark, and the presence of W can be demonstrated in a court of low (that is, W could say, "This is the Property of X, and was licensed to Y on date Z");
3. The watermark W is not readable from T' without knowledge of the secret key that was used to introduce W;
4. For someone who knows the secret key, W can be obtained from T' without knowledge of T;
5. Unless someone knows the secret key, W is impossible to remove from T' without significantly changing the meaning of T';
6. The process by which W is introduced into T to obtain T' is not secret, rather, it is the secret key that makes the scheme secure;

The following sections describe the above approach in detail.

3.1 Embedding Watermark

Watermark embedding involves five steps - term segmentation and tagging, secret key generation, candidate selection, semantic substitution and ciphertext generation. Figure 2 presents the process of watermark embedding.

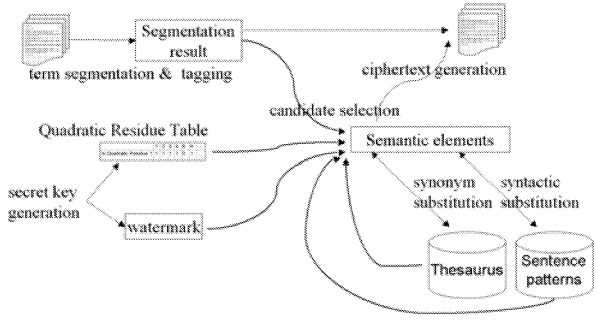

Fig. 2. Steps involved in watermark embedding

3.1.1 Term Segmentation and Tagging

Term segmentation and tagging are essential to semantic processing especially for the proposed approach. In this study, the terms are segmented and tagged using AutoTag - a segmentation and tagging tool for Chinese. AutoTag was developed by the Chinese Knowledge Information Processing group (CKIP), which is a Chinese term segmentation system designed by Academia Sinica. AutoTag has the richest lexical dictionary in Chinese.

3.1.2 Secret Key Generation

In this study, the secret key has two primary components - watermark and user quadratic residue key. First, the watermark is translated into a bit string to create the secret encoding key. Unicode, ASCII, ANSI or other techniques can be used to conduct this translation. This study uses Unicode. Assuming that the watermark is 『ACT』, and its bit string is like 『00101』, then the quadratic residue key is a prime candidate for producing a quadratic residue table. This table then can be used to identify which number is residue, or non-residue between 1 and the quadratic residue key. Here residue indicates 1, and non-residue indicates 0. Subsequently, this table is used to help in term selection, as mentioned later. Table 1 illustrates the quadratic residue table of the prime number 10007.

Table 1. Quadratic Residue Table of prime number 10007

	1	2	3	4	5	6	…	10007
Is Quadratic Residue ?	1	1	1	1	0	1	…	0

3.1.3 Candidate Selection

Candidate selection involves selecting candidate elements for syntactic or synonym substitution for watermark embedding. In syntactic substitution, the sentences matching the defined sentence patterns are selected. However, synonym substitution must consider additional factors. Not all terms in the text can be used for watermark embedding, and unsuitable terms should be filtered out. The filtering rules generalized here are illustrated in Fig. 3:

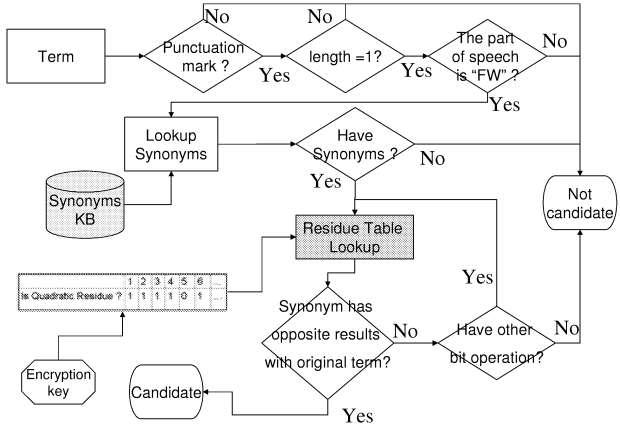

Fig. 3. Term selection flow

1. Punctuation mark: Punctuation marks in the text are fixed and difficult to tune. Punctuation marks thus are ignored for the present purposes.
2. The length of the term segmented is 1 :In AutoTag, terms with a length is 1 mostly comprise prepositions(介詞, P), DE(e.g. 的), or pronouns(代名詞, Nh). These terms are either difficult or impossible to substitute. Therefore, these terms also are skipped.
3. The part of speech is "FW" in AutoTag: In AutoTag, the part of speech of the term "FW" means foreign language, such as English, Japanese, that do not belong to Chinese. This study ignores the issue of how to treat such terms.
4. The term has no synonym: The approach proposed here uses synonym substitution to embed the watermark in the text. Consequently, if the term has no synonyms, it will not be embedded in the watermark. The term thus must be truncated.
5. Although the term does have synonyms, no opposite results appear by looking up the quadratic residue table after the following bit operations.

This approach uses bit matching to embed a watermark in the text. Therefore, the opposite result is needed. Synonyms exist which can be substituted when bit matching fails. This study uses the following bit operations： XOR, AND, OR, +, -, /. Assuming that a term in Chinese "分配"(dispatch), can be divided into two characters and their bit strings - 『分：10011』和『配：10101』. After performing the AND bit operation, "分配"(dispatch) is expressed by『10001』and the decimal number

『17』. Then, the quadratic residue key (e.g. 10007) is taken to modify this decimal number and the remainder are used to look up the quadratic residue table. This process obtains result 1. The quadratic residue table results of these three terms are listed in Table 2 below:

Table 2. Some quadratic residue table results of three terms

Term	Synonym	Residue：1	Non-Residue：0
分配(dispatch, Verb)	發給(send), 分發(deliver)	分配(dispatch), 發給(send)	分發(deliver)
賺取(earn, Verb)	創利(make), 贏利(profit), 盈利(gain), 創收(obtain)	盈利(gain), 創收(obtain)	賺取(earn), 贏利(profit), 創利(make)
情形(situation, Noun)	情況(circumstance), 狀況(status), 狀態(state), 條件(condition)	情形(situation), 狀況(status), 條件(condition)	情況(circumstance), 狀態(state)

3.1.4 Semantic Substitution

Semantic substitution involves two main operations --- syntactic substitution and synonym substitution. These operations can be used to embed the watermark in the text. First, this study performs embedding by syntactic substitution. The candidate sentences are selected based on the substitution rules gathered here. Tree structures then are used to express those sentences. Figure 4 is an example of the tree of sentence "月光和角落裡的貓"(Moonlight, and a cat in the corner).

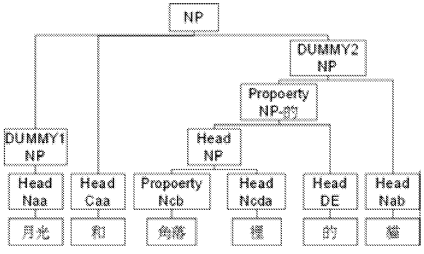

Fig. 4. Tree structure of a Chinese sentence

A Chinese Parser was used to construct the tree structure of a Chinese sentence. With the emergence of large treebanks, supervised statistical English parsers are achieving promising results. Penn Chinese Treebank [9] has an average sentence length of 30 words and presents a rich source of Chinese treebanks. Extraction of Probabilistic Context Free Grammar (PCFG) from a treebank is straightforward [5]. This study adopted the Chart parsing algorithm in[3], that depicts each symbol position as a unique role using CFG grammar rules, This study started to consider that when a category plays different roles, the different probability distributions of its expanding rules will embody the subtlety of phrasing preferences.

Every node in the tree structure represents a part of speech. Moreover, every part of speech is assigned a unique score for calculating the Depth First Search (DFS) and Breadth First Search (BFS) scores.

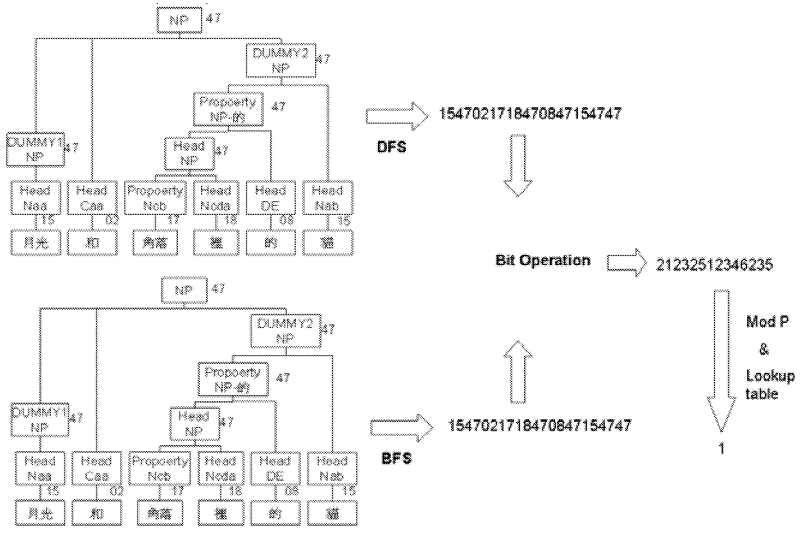

Fig. 5. Look up the residue value of the origin sentence

A residue value then can be obtained by a bit and mod operation. The quadratic residue key is used to perform the mod operation to look up quadratic residue table (see Fig. 5). Syntactic substitution is not performed when the residue values of the synonymous and origin sentences are the same. Otherwise, syntactic substitution is performed (see Fig. 6). After completing all syntactic substitutions, the watermark is embedded by synonym substitution.

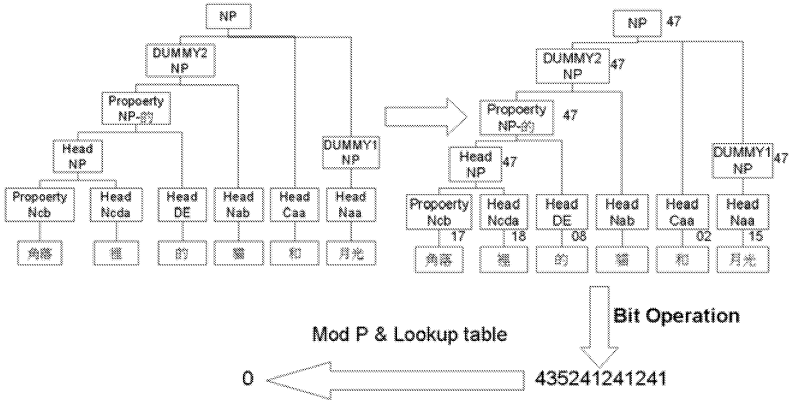

Fig. 6. Look up the residue value of the substitute sentence

In synonym substitution, the bit string of the watermark is taken to embed it in the text in order with one bit being embedded in each embedding. When 0 is embedded in the text, and the quadratic residue value in the current term is 1, then the substitution

is performed. Otherwise, no substitution is required. Additionally, this study also considers semantics following substitution. Ciphertext should be similar to plaintext semantically. Thus, the embedding is said to be successful. The synonym selection rules in this study are as follows:

1. Opposite results in the quadratic residue table : A term may have numerous synonyms. If term substitution is required, then terms with opposite results need to be selected in the quadratic residue table.
2. The part of speech of the term should be exactly the same as the original term: A term may indicate numerous parts of speech in different scenarios. Therefore, the part of speech of the term should be considered in substitution.
3. Common usage: One or more synonyms with the same the part of speech. Thus, the corpuses should be used to select one synonym with higher common usage.

After selecting the synonyms through the above three rules, the appropriate synonym must be selected to preserve the same meaning, and can embed maxima watermark bits. All synonyms appear in the corpus, and thus can maintain their meaning unchanged. The Binary Tree based Synonyms Selection algorithm was proposed, which can select the synonym embed maxima watermark bits. The BTSS algorithm involves tree steps:

1. Translate all words in all synonyms to bit strings.
2. Construct the binary tree.
   (1) Group all synonyms together.
   (2) Divide all synonyms to two sub-groups by bit operation and Quadratic Residue Table lookup, as described in section 3.1.3. The bit operation is applied sequentially. Previously used bit operations cannot be reused.
   (3) Each sub-group should be divided into two sub-groups using step ii, until a sub-group has one synonym or no bit operations can be used.
3. Selecting the appropriate synonym embeds maxima watermark bits based on the watermark bit strings.

According to the above example, the term ”賺取”(earn) has four synonyms {“創利”(make), ”贏利”(profit), ”盈利”(gain), ”創收”(obtain)}. For the step1, all of the words in all synonyms, including the term ”賺取”, are translated to bit strings as listed in Table 3.

Table 3. Bit strings for each synonym

Term	Translated to bit strings for each words in term
賺取(earn)	『賺：10011』, 『取：01001』
創利(make)	『創：11010』, 『利：01101』
贏利(profit)	『贏：00111』, 『利：01101』
盈利(gain)	『盈：10111』, 『利：01101』
創收(obtain)	『創：11010』, 『收：01110』

Step2 first used the bit operation OR to calculate each synonym and consult the Quadratic Residue Table. Table 4 lists the result.

Table 4. Quadratic Residue for each synonym

Term	Translated to bit strings for each words in term	Bit Operation OR	Quadratic Residue
賺取(earn)	『賺：10011』, 『取：01001』	『賺取：11011』	0
創利(make)	『創：11010』, 『利：01101』	『創利：11111』	0
贏利(profit)	『贏：00111』, 『利：01101』	『贏利：01111』	0
盈利(gain)	『盈：00101』, 『利：01101』	『盈利：01101』	1
創收(obtain)	『創：11010』, 『收：01110』	『創收：11110』	1

Table 4 reveals that the terms "賺取"(earn), "創利"(make) and "贏利"(profit) are grouped together. Additionally, the terms "盈利"(gain) and "創收"(obtain) also are grouped together. The bit operation add(+) is used for the "盈利"(gain) and "創收"(obtain) group, and Table 5 shows the result.

Table 5. Quadratic Residue for each synonym in sub-group {盈利(gain), 創收(obtain)}

Term	Translated to bit strings for each words in Term	Bit Operation add	Quadratic Residue
盈利(gain)	『盈：00101』, 『利：01101』	『盈利：10010』	1
創收(obtain)	『創：11010』, 『收：01110』	『創收：01000』	0

From Table 5, the terms "盈利"(gain) and "創收"(obtain) are divided into two sub-groups. Figure 7 illustrates the Final Binary tree.

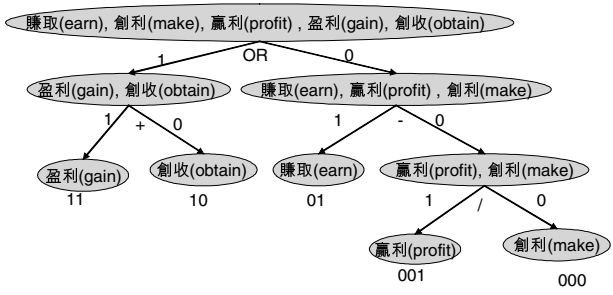

Fig. 7. The binary tree of synonyms

In step 3, the appropriate synonym embeds maxima watermark bits according to the watermark bit strings. If the watermark bit strings are 001, the term贏利(profit) is the appropriate synonym. Moreover, if the watermark bit strings are 11, then the term盈利(gain) is the appropriate synonym. The BTSS algorithm proposed here overcomes the drawback of one synonym only being able to encode one watermark bit. Because the BTSS algorithm is based on the binary tree, the average bits which can be embedded in synonyms are log N, where N denotes the number of synonyms.

3.1.5 Ciphertext Generation

The ciphertext is generated after the proceeding the above steps. Moreover, the author can spread the ciphertext in public. When unauthorized usage occurs the watermark can be extracted to demonstrate ownership. Section 3.2 describes how to extract the watermark from the text.

3.2 Watermark Extraction

The rapid development of a network multimedia environment has enabled digital data to be distributed faster and more easily than ever before. Duplicating digital data thus also has become easier than previously. The preceding section described how to embed the watermark in a text. Meanwhile, this section explains how to extract a watermark from a text when unauthorized usage is suspected. Except for the lack of synonym substitution, the other steps involved are similar. Figure 8 presents the process of watermark extraction.

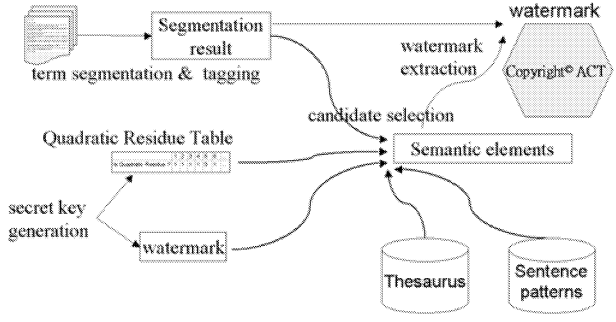

Fig. 8. Watermark extraction process

The first three steps, namely term segmentation and tagging, secret key generation and term selection, are the same as for watermark embedding. Following term selection, residue values can be looked up from the quadratic residue tale in the last step. Those residue values then can be assembled into a bit string. Finally, passed through transformation between the bit string and Unicode to exact the watermark.

4 Experiments

The training data are 200 related report documents dealing with the computer industry from III (Institute for Information Industry) and MIC (Market Intelligence Center). The average terms count of these documents is 1000. Moreover, the average sentences count of these documents is 150. Additionally, the average number of terms selected for hiding the watermark bit string is 120. Furthermore, the average number of sentences selected for hiding the watermark bit string is ten. Finally, the average number of synonyms for each term is three.

As described in section 3.1.4, the BTSS algorithm can enhance the numbers of watermark bits from one bit to log n bits per synonym. The watermark can be hidden in the text as often as possible. The numbers of times that watermark can be hidden depend on the length of the text. In the standard version, which does not use the BTSS algorithm, a relationship of ca. 1:10 exists between the encoded information and the protected text (e.g. for every ten terms of text one bit of watermark can be hidden). The version, which uses the BTSS algorithm, creates a relationship of ca. 1:6.6 between the encoded information and the protected text. Table 6 lists the experimental result.

Table 6. The experimental result

	Without BTSS	With BTSS
The average number of sentences	150	150
The average number of sentences selected for embedding watermark bit string	10	10
The average number of terms	1000	1000
The average number of terms selected for embedding watermark bit string	120	120
The average number of synonyms for one terms	3	3
Average bits hidden in one synonyms	1	1.5
Relationship of ca.	1:10	1:6.6

Two kinds of tests are conducted to measure the perceptual transparency. Blind test, where human subjects are presented with protected text without original text. Non-Blind test, where human subjects are presented with protected text with original text. The test data are fifty documents and ten people are invited to measure the perceptual transparency. The PSNR is used to measure the perceptual transparency. $PSNR = \frac{\sum U_i}{\sum S_i}$, where S_i is the total number of sentences for each document. And U_i is the total number of unsuitable sentences for each document. The PSNR in Blind test is 0.03 and in Non-Blind test is 0.0699. Table 6 lists the experimental result of transparency.

Table 7. The experimental result of transparency

	Blind test	Non-Blind test
The average number of sentences	150	150
The total number of sentences	75650	75650
The total number of unsuitable sentences	2270	5290
The average number of terms selected for embedding watermark bit string	0.0300	0.0699

5 Conclusions and Future Works

This study proposed a synonym-based watermarking algorithm, which is context rather than format based. The synonym-based watermarking is suitable for Chinese textual messages. The algorithm was designed to select appropriate candidate terms from textual messages for use in the embedding watermark. The binary tree encoding methodology of the synonym-based watermarking algorithm can select the suitable synonym, which embeds the maximum watermarking bits. The average watermarking lengths that can be encoded in one term are log n bits, where n denotes the total number of synonyms of the term. The encoding/decoding algorithm of the encryption key based watermark also is proposed, and enables higher watermark resilience. Besides extending the thesaurus to improve the reliability of synonym substitution, future efforts should explore the Syntactic and Semantic approaches that not only maintain consistent context semantics but also enable increased watermark resilience. In the Syntactic approach, watermark bits are hidden in the structure of sentences. The converting technology for the same meaningful sentence, including the co-reference, zero anaphora and grammar approaches, is used to deal with the problem of watermark encoding.

Acknowledgement. This research was supported by the III Innovative and Prospective Technologies Project of Institute for Information Industry and sponsored by MOEA, ROC

References

1. Brassil, J., Low, S., Maxemchuk, N., and O' Gorman, L., Electronic marking and identification techniques to discourage document copying. Proceedings of IEEE INFOCOM '94, 1994 3, pp. 1278–1287.
2. Brassil, J., Low, S., Maxemchuk, N., and O' Gorman, L. Hiding information in document images. Proceedings of the 29th Annual Conference on Information Sciences and Systems, 1995, pp. 482–489.
3. Charniak, E. 1997. Statistical parsing with a context-free grammar and word statistics. In Proceedings of NCAI-1997, pp 598–603.
4. Hardy, G. H. and Wright, E. M. "Quadratic Residues." §6.5 in An Introduction to the Theory of Numbers, 5th ed. Oxford, England: Clarendon Press, pp. 67–68, 1979.
5. Johnson, M. 1998. The effect of alternative tree representations on tree bank grammars. In Proceedings of the Joint Conference on New methods in Language Processing and Computational Natural Language Learning (NeMLaP3/CoNLL'98), pp 39–48.
6. Low, S., Maxemchuk, N., Brassil, J., and O' Gorman, L., Document marking and identification using both line and word shifting. Proceedings of IEEE INFOCOM '95, 1995.
7. M. Atallah, V. Raskin, C. F. Hempelmann, M. Karahan, R. Sion, K. E. Triezenberg, U. Topkara, "Natural Language Watermarking and Tamperproofing", Proc. of the Information Hiding Workshop IHW 2002, Lecture Notes in Computer Sciences, Springer Verlag (LNCS)
8. Mikhail J. Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. "Natural Language Watermarking: Design, Analysis, and Proof-of-Concept Implementation" published in the Proceedings of the 4th International Information Hiding Workshop, Pittsburgh, Pennsylvania, April 25–27, 2001.
9. Xia, Fei; Palmer, Martha; Xue, Nianwen; Okurowski, Mary Ellen; Kovarik, John; Chiou, Fu-Dong; Kroch, Tony and Marcus, Mitch (2000) Developing Guidelines and Ensuring Consistency for Chinese Text Annotation in Second International Conference on Language Resources and Evaluation (LREC-2000) pp. 3–10.