Chinese Text Zero-Watermark Based on Three-dimensional Space Model

Yingjie Meng
School of Information Science & Engineering, Lanzhou University, Lanzhou, China
Email: mengyj@lzu.edu.cn

Liming Gao
School of Information Science & Engineering, Lanzhou University, Lanzhou, China
Email: gaolm09@lzu.edu.cn

Mingwen Liu
School of Information Science & Engineering, Lanzhou University, Lanzhou, China
Email: liumw11@lzu.edu.cn

Lixin Bai
School of Information Science & Engineering, Lanzhou University, Lanzhou, China
Email: bailx11@lzu.edu.cn

***Abstract*—Aiming at the limitation of existing Zero-Watermark algorithm in the copyright protection, a Chinese text Zero-Watermark scenario based on three-dimensional space model was proposed. In the scenario, the three-dimensional model was constructed by using the features of high frequency words, sentence length and its weights. Afterwards, the algorithm generates a watermark based on the text abstract-set which can be extracted later by comparing the distance of each sentence point with the centre of model. In case of any copyright conflict, the extraction algorithm extracts the abstract-set from the copyright disputes text again, and by calculating the semantic distance to define original copyright owner. Furthermore, the effectiveness and feasibility of the proposed algorithm was proved with algorithm simulation. The simulation results also showed that the algorithm had good robustness, especially for syntactic transfer, synonyms replacement and shear attacks occurring randomly.**

***Index Terms*—Digital text; Copyrights Protection; Zero-watermark; Three-Dimensional Space**

I. INTRODUCTION

With the widespread use of Internet and other communication technologies in recent years, the network delivery of novels, journals and documents in digital way is more and more popular. In addition to getting the benefits of information exchange, the digital community is confronted with authentication, forgery, and copyright protection issues. Most of this kind information can be categorized or directly converted to text formatting,

Causing them to be copied easily. Thus, how to protect the information effectively has attracted people's close attention, and text digital watermark technology are increasingly used to meet this challenge and identify the copyright.

Currently, research on the text digital watermark can be divided into two main categories. The traditional one belongs to embedded watermark in text image. Brassil, et al. first proposed a few text watermarking methods by changing the editorial format(shifting the words and sentences upwards and downwards) to embed watermark bits [1,2]. Later Maxemchuk and Low, et al [3,4,5] evaluated and analyzed the efficiency of above mentioned methods. Atallah and some others proposed syntactic watermark by using syntactic structure of text [6,7]. Another embed algorithm based on text meaning representation (TMR) string has also been proposed [8]. Although these treatments can protect the text, the watermark is embedded in the text will affect the value of original data. In this paper, we proposed the other algorithm called zero-watermark. A typical zero-watermark divided in two processes: embedding algorithm and extraction algorithm. Watermark embedding is done by original author and extraction done later by Certifying Authority (CA) to prove ownership. This scheme does not change the feature of text, but utilize the characters of original text to construct watermark so that the value of original one can be well protected.

The existing Chinese text zero-watermarking algorithm mainly based on key words and phrases extraction from the text to construct the watermark, such as Si Lei, et al [9] proposed a Chinese texts zero-watermarking algorithm based on key-words, and then Guo Tao, et al [10] using sentence's entropy to construct Chinese texts zero-watermarking algorithm. But these algorithms are based on two-dimensional space model of words-level, and words and phrases are just the basic components of the text, although it better reflects the local characteristics of text, it ignores the complete

doi:10.4304/jcp.7.8.2063-2070

semantic and structural information, and it can not outstandingly express theme characteristic of the whole text content.

Based on the description of above, this article introduces three-dimensional space model to select the sentence-set that reflect the main feature of text, then calculate the similarity of the watermark by semantic distance, and on the basis of this, a text digital watermarking algorithm based on three-dimensional space model was proposed. The result of experiment shows that our scenario has greater ability of anti-attack and more robustness than other watermark scenarios.

This paper is organized as follows: The definition of the algorithm is given in section 2. Section 3 described the scheme design of watermark. The proposed embedding and extraction algorithm are illustrated in detail in section 4. Section 5 presents the experimental result on texts with random attacks. Performance of the proposed algorithm is analyzed with multiple attack volumes. The last section concludes the along with direction for future research.

II. DEFINITION

In order to construct and describe the watermarking scheme, we represent the *T* as the Chinese text to be protected and we will give some definitions and stipulations to illustrate the watermark scenario.

Definition 1 Word-set *W*, a set which comprises *n* words after trim punctuations, word segmentation and delete the stop-words and other useless information. *Wi*($1 \le i \le n$) is the ith word of *W*, *Wi.data*, *Wi.freg*, *Wi.lengh*, *Wi.count1*, *Wi.count2*, *Wi.count3*, *Wi.weight* respectively mean the value of *Wi*, word-frequency, word length, a ratio of the times of this word appears in the title, in the first-sentences, at the end of the paragraphs and weights.

The weight of word *Wi.weigh*t, it reflects the importance of word *Wi* in the text. Based on the idea of [11], we constructed formulas as described in (1).

$$Wi.weight = a \times \frac{Wi.freg}{1+Wi.freg} + b \times \frac{Wi.lengh}{1+Wi.lengh} + c \times \frac{10 \times (Wi.count1 \times 5 + Wi.count2 \times 3 + Wi.count3 \times 2)}{n} \quad (1)$$

Among them, *a*, *b* and *c* are the proportionality factors, which used to indicated various factors in the proportion of wighted formula, and $a+b+c=1$.

Definition 2 Sentence-set *S*, a set which comprises m sentences after trim punctuations, *Sj*($1 \le j \le m$) is the jth sentences of *T*, *Sj.data*, *Sj.lengh*, *Sj.num*, *Sj.weight*, Sj.position respectively mean the value of *Si*, the number of words, the number of high frequency words, weights, and importance of the position of the sentence, among them, the value of Sj.position is :

$$S_j.position = 1 - \frac{j}{m} \left(1 \le j \le m\right)$$

The weights of sentences *Sj.position*, it reflects *Sj* status of the entire text, which is the weight, location, length and other information of words contained in *Sj*.

The calculation formula as (2)

$$Sj.weight = \alpha \frac{\sum_{Wi \in Sj} Wi.weight}{\max_{Wi \in Sj} \{ Wi.weight \}} + \beta \frac{Sj.lengh}{1+Sj.lengh} + \gamma Sj.position \quad (2)$$

Among them, α, β and γ are the proportionality factors, which used to indicated various factors in the proportion of wighted formula, and $\alpha+\beta+\gamma=1$

Definition 3 π is threshold of high frequency words, namely the number of high frequency words selected from Text set.

Definition 4 Text focus $O(k_x, k_y, k_z)$, is the centre of the text three-dimensional model, The calculation formula as (3)

$$\begin{cases} k_x = \frac{1}{m}\sum_{i=1}^{m} S_{jx} \\ k_y = \frac{1}{m}\sum_{i=1}^{m} S_{jy} \\ k_z = \frac{1}{m}\sum_{i=1}^{m} S_{jz} \end{cases} \quad (3)$$

Definition 5 ε is threshold of summary, namely the number of summary selected from Sentence-set.

Definition 6 Abstract-set *Abs*, is the set that containing ε summaries.

Definition 7 semantic distance distance(*Si*,*Sj*), show the sentence *Si* and *Sj* semantic distance, which is the calculation of two sentences on the content similarity, any two sentence Si and *Sj* semantic distance is defined as follows:

$$distance(S_i, S_j) = 1 - \frac{samewc(S_i, S_j)}{len(S_i) + len(S_j)} \quad (4)$$

Among them, samewc(*Si*,*Sj*) is the same number of words in the two sentences. We count the less number when a word appear in Si is not the same as the number of *Sj*. Obviously, the sentence semantic distance between the range located in the interval [0,1], the greater the semantic distance, indicating that the difference in the content is larger.

Definition 8 ω is threshold of text similarity, namely the similar degree of text based on the semantic distance.

III. CHINESE TEXT ZERO-WATERMARK SCHEME DESIGN

A. The Three-Dimensional Space Model of Text

Text is the orderly arrangement of sentences, and sentences are composed of words. The text were once described by words or phases to describe, but its size is too small and text semantic expression is low. In this

paper, therefore, we use sentence-level description, which will make text semantic understanding more detailed and accurate, at the same time, can better grasp text important characteristics.

Using the length of sentence and whether it contains high frequency words of text as the two basic attributes of the text, On that basis, we can construct the three-dimensional space model *Sj(Xj,Yj,Zj)*of text T. As shown in Fig 1, the value of Y axis follow: No high frequency words in Sj, we take 0, while the number less than π, take 1, otherwise take 2.

According to the figure1, mapping all sentences of the text in this model can produce three-dimensional space model of entire text. After the structure, we can calculate the centre of this three-dimensional space model through the formula (3), which is the text focus, then calculate the distance of each sentence point *Sj(Xj,Yj,Zj)* to the centre of model, show as Disj ,finally generate distance-set Dis and select first ε sentences as the summaries from this set.

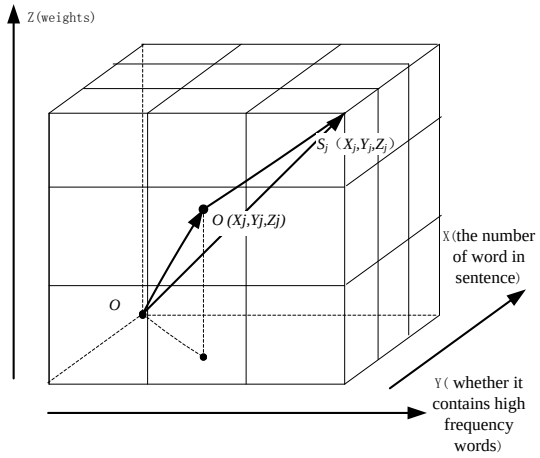

Figure 1. Three-dimensional space model of S*j*

B. Structure of Test Zero-Watermarking

In order to reduce the amount of work and complexity of whole scenario, pre-process work for the text should be done firstly. Primary work of this part is to divide the text into sentences and words, so that we will get word-set W, and this specific method has been relatively mature, which was mentioned as in [12]. Considering our research focuses is the extraction and examination of zero-watermarking. So these work and its details are not discussed here.

Chinese text zero-watermarking in three-dimensional space model consists of two main components: the constructing and the detecting

1) structure of zero-watermarking

The structuring of zero-watermarking based on word-set W. Firstly, we constructed the three-dimensional space model of text T. Then calculate the distance of each sentence point to the centre of model, in order to selecting the abstract-set of text, and this set is used as watermark of text. Finally, we can introduce a third-party authoritative organization like Certifying Authority (CA) to register watermark. According to the above parts, concrete construction process can be designed as Fig 2

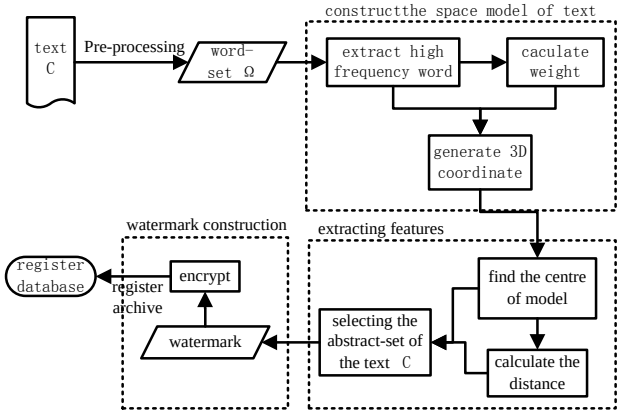

Figure 2. Construction process of text zero-watermark

2) detecting of zero-watermarking

Watermark detecting is to judge the watermark when copyright dispute appeared. Work of this part is similar to the constructing watermark, which also need to construct disputed-text's watermark again, then calculate the similarity of disputed-text's watermark and stored watermark by detecting algorithm in order to determine copyright ownership. The overview of watermark detecting is shown in Fig 3.

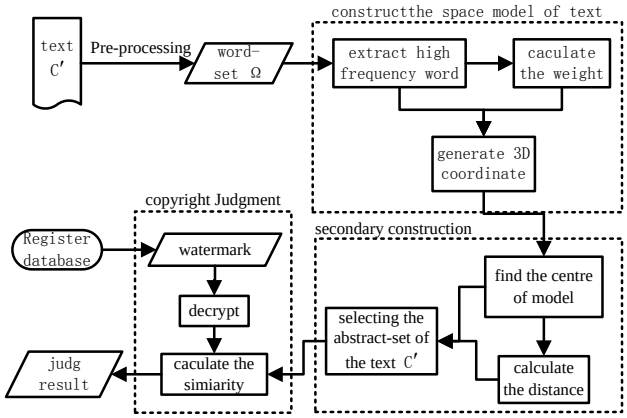

Figure 3. Detect process of text zero-watermark

IV. Algorithm of Watermark Construction

Based on the description of above and Fig 2, the construct algorithm of text zero-watermark can be designed as followings. The algorithm conducts on the word-set W, The procedure of constructing watermark comprises three algorithms, including getting three-dimensional coordinates of sentences, extracting the summaries and generating watermark.

1) algorithm of getting three-dimensional coordinates of sentences

In the whole construction process of watermark, we can get three-dimensional coordinates of sentences according to the figure 1, the process can be divided into two phases:

1) Generating high-frequency words set;
2) Coordinate generation

Generating high-frequency words set is the most basic part of watermark construction, which is got by calculate the weights of each word in word-set. The algorithm Keyword-Select executes as following.

PROCEDURE *Keyword-Select* (*W*, *π*, *a*, *b*, *c*, *KP*)
Input: *T'* word-set *W*[1…*m*], threshold *π*, proportionality factors *a*, *b*, *c*;
Output: high-frequency words set *KP*[1…*π*]
BEGIN
FOR *i*←1 TO *n* DO
W[*i*].*weight*←*a***W*[*i*].*freg*/(1+*W*[*i*].*freg*)+*b***W*[*i*].*lengh*/(1+*W*[*i*].*lengh*)+*c**10*(*W[i]*.*count*1*5
+*W*[*i*].*count*2*3+ *W*[*i*].*count*3*2)/*n*;
Call Sort(*W*); *//order W by W[i].weight descending*
FOR *i*←1 TO*π* DO
KP[*i*] ←*W* [*i*]
END

After the high-frequency words set KP was generated, we can construct three-dimensional space model of text, in this process, Let S[j].lengh, S[j].num represent the number of words in Sj and the number of high-frequency words. The X, Y and Z axis which stored by record array TS[1…m]. The data item TS[j].X, TS[j].Y, TS[j].Z represent the X, Y and Z axis of sentences Sj respectively. The specific algorithm is shown as following:

PROCEDURE Coordinate-Create (*S*, *W*, *π*, *a*, *b*, *c*, α, *β*, *γ*)
Input: *T'* word-set *W*[1…*n*], threshold *π*, proportionality factors α, *β*, *γ*, high-frequency words set *KP*
Output: a array stored 3d coordinates of sentences *TS*[1…*m*]
BEGIN
FOR *j*←1 TO *m* DO
[*i*←1; *S*[*j*]. *lengh*←0; *S*[*j*].*num*←0;
WHILE (*i*≤*n*) DO
[IF (*W*[*i*].*data*∈*S*[*j*].*data*) THEN
//Counting the number of words in S_j
S[*j*].*lengh*←*S*[*j*].*num*+1;
IF (*W*[*i*].*data*∈*KP* [*i*].*data* THEN
//Counting the high-frequency *words in* S_j
S[*j*].*num*←*S*[*j*].*num*+1;
i←*i*+1]]
FOR *j*←1 TO *m* DO
[*//generate X axis*
TS[*j*].*X*←*S*[*j*].*lengh*;
//generate Y axis
CASE
S[*j*].*num*=0: *TS*[*j*].*Y*←0;
S[*j*].*num*=*π*: *TS*[*j*].*Y*←2;
ELSE TS[j].Y←1;
ENDCASE
//generate Z axis
sum←0; *max*←0; *i*←1;
WHILE (*i*≤*n*)and (*W*[*i*].*data*∈*S*[*j*].*data*) DO
[*//calculate sum*
sum←*sum*+ *W*[*i*].*weight*;
IF (*W*[*i*].*weight* >*max*) THEN
// calculate max
max←*W*[*i*].*weight*;
// calculate max
i←*i*+*1*;]
S[*j*].*position*←2-*j*/*m* ;
// calculate the weight of S_j
S[*j*].*weight*←*α***sum*/*max*+*β***S*[*j*].*lengh*/(1 + *S*[*j*].*lengh*) +*γ***S*[*j*].*position*
TS[*j*].*Z*←*S*[*j*].*weight*]
END

2) algorithm of extracting sumarries

After getting the three-dimensional coordinates of sentences, we calculate the centre of the three-dimensional space model, that is $O(k_x, k_y, k_z)$, then selecting the abstract-set Abs through calculating the distance between the each sentence point and $O(k_x, k_y, k_z)$, the process can be divided into two phases: Generating the distance-set Dis;Abstract-set Abs getting.

PROCEDURE *abstract-create*(*S*, *TS*[1…*m*], *ε*)
Input: a array stored 3d coordinates of sentences *TS*[1…*m*]
Output: the *Abstract-set*
BEGIN
// calculate the text focus

$$k_x \leftarrow \frac{1}{m}\sum_{i=1}^{m} S_{jx};$$

$$k_y \leftarrow \frac{1}{m}\sum_{i=1}^{m} S_{jy};$$

$$k_z \leftarrow \frac{1}{m}\sum_{i=1}^{m} S_{jz};$$

// generate distance-set Dis
FOR *i*←1 TO *n* DO
[$Dis_i.value \leftarrow \sqrt{(TS[i].X - k_x)^2 + (TS[i].Y - k_y)^2 + (TS[i].Z - k_z)^2}$;
$Dis_i.index \leftarrow i$;]
//order Dis by dis.value ascending
Call sort(*Dis*)
FOR *i*←1 TO ε Do
Abs[i] ←S[Dis_i.index];
END

3) algorithm of getting watermark

Based on the three-dimensional coordinates of sentences generation and abstract-set extraction, we can construct the complete watermark generating algorithm *structuring*. The basic function of the algorithm is: By

calling *Coordinate-Create* and *abstract-create* through the word-set *W*[1…*n*] and sentence-set *S* [1…*m*] generate the abstract-set *Abs*, in order to improve the security of the watermark, *TS*[1…*m*] is encrypted with SHA-1 encryption as in paper [13]. Encrypted watermarking information stored in the one-dimensional array *List*[1…*m*], and the algorithm *Structuring* is shown as following:

```
PROCEDURE Structuring (S, W, List, Key)
Input: sentence set S[1…m], word-set W[1…n], Key
Output: watermark List[1…m]
BEGIN
    // generate high-frequency words set
    CALL Keyword-Select (W, π, a, b, c);
    // generate 3d coordinates of sentences
    CALL Coordinate-Create (S, W, π, a, b, c, α, β, γ);
    //generate abstract-set
    CALL abstract-create (S, TS[1…m], ε);
     FOR j←1 TO ε DO
      // use SHA-1 encryption
      List[j] ←SHA-1(Abs[j], Key);
END
```

B. Watermark Detection

The watermark scenario is on the base of zero-watermark, so before detection for disputed-text T′, we still need to construct the watermark Abs′ again, and the procedure is similar to that of construction. The semantic distance of Abs′ and archived Abs are calculated in order to determine copyright ownership.

1) calculate similarity

The main work of similarity calculation is to calculate the similarity of disputed-text and the original one, and estimate whether the two watermarks are same to each other, which use the formula (4), and the algorithm Calculate-Similarity is shown as following:

```
PROCEDURE Calculate-Similarity (Abs, Abs′)
Input: abstract-set Abs, Abs ′
Output: Similarity
BEGIN
    Similarity ← 0 ;
    For i←1 TO ε DO
    [ len1= len(Abs_i);// get length of sentence
      len2= len(Abs′_i);
      samewc←0;
      IF len1 >= len2 THEN
         For j←1 TO len2 DO
         [ //if Abs′_i has word in Abs_i samewc add 1
            IF Abs′_i[j] in Abs_i THEN
            samewc++;[
      ELSE
         For j←1 TO len1 DO
         [ //if Abs_i has word in Abs′_i samewc add 1
            IF Abs_i[j] in Abs′_i THEN
            samewc++;]
      distance ← [1 − samewc / (len1 + len2)];
      Similarity ← Similarity + distance / ε ;]
END
```

2) watermark detection algorithm

Based on the similarity calculations mentioned above, the main process of detection algorithm detecting is described below

```
PROCEDURE Detecting (S′, W′, List, ω, Key)

Input: disputed-text, S′, W′, List, threshold ω, Key
Output: whether watermark can be detected
BEGIN
    // generate high-frequency words set
    KP ′← CALL Keyword-Select (W′, π, a, b, c);
    // generate 3d coordinates of sentences
    TS′←CALL Coordinate-Create (S′, W′, π, α, β, γ, KP′);
    //generate abstract-set
    Abs′←CALL abstract-create (S′, TS′[1…m], ε);
    //get Abs from Watermark
    List←Watermark;
       FOR j←1 TO ε DO
          // use SHA-1 decryption
          Abs[j]←SHA-1^-1(List [j], Key);
    //calculate the similarity
    Similarity←Calculate-Similarity (Abs, Abs′)
       IF Similarity>ω THEN
            Writeln(' watermark exits.')
       ELSE
            Writeln('no watermark');
END
```

V. Experiment and Performance Analysis

In order to verify the feasibility and availability of the proposed model, we conducted a simulation experiment. The experimental environment is listed as below: CPU: Intel Core2 Duo E4600/2.80GHz, RAM: 1.0GB, Windows XP; language Visual C + + 6.0. Subjects: a word document named "The life of blue sky". The total word is 1382. The process for segmenting the text uses the algorithm of [14]. Relevant parameters for experimental are as follows: threshold of keywords π is set to 3, the similarity threshold ω set to 0.6, Based on the experiments in [11,12], we set proportionality factor *a* as1.5, *b* as 1.0, *c* as 0.8; *α* as 0.3, *β* as 0.3 and *γ* as 0.4. Firstly we discuss the threshold in this algorithm, then verify the robustness of watermark after implementing three most common attacks: syntactic transfer, synonym replacement, and modification, in order to test the performance of the algorithm.

A. The Impact of Summary Extraction Threshold on Watermark

The analysis about the watermark construction show that summary extraction threshold ε determine the robustness of watermark to a large extent. So, this experiment mainly analysis the influence of the threshold ε to the algorithm. The experiment use word document named "The life of blue sky". The total word is 1382.

After implementing three most common attacks: syntactic transfer, synonym replacement, and cut on the word document, the influence of the summary extraction threshold ε selecting to the watermark generation as shown in figure 4 shows. The horizontal axis of the figure

represents the drawing of ε, the vertical axis represents similarity of text, similarity threshold ω set to 0.6.

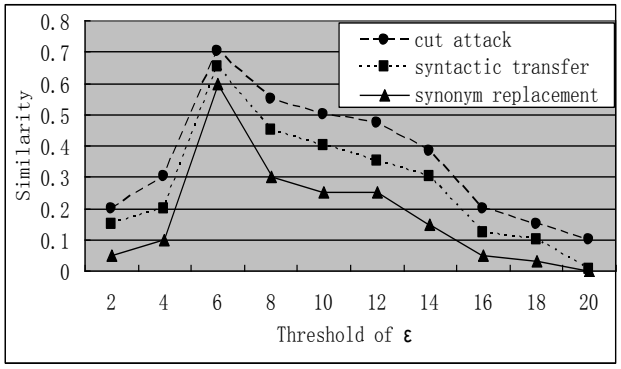

Figure 4. The Influence of summary extraction threshold on watermark

The three curves show that when text under the three attacks mentioned in the figure 4, the watermark detection level is rising with the increase of summary extraction threshold ε, when setting the summary extraction threshold ε as 6, watermark detection can achieve the best degree, higher than our default text similarity threshold 0.6. The watermark detection level presents the downward trend after the threshold ε increased further. So we set the topic paragraph to 6.

B. Syntactic Transfer Attack

Syntactic transfer is one of the most frequent text attacks, such as the conversion between the two sentences turns active into passive, bring words to the preposition and tenses changes. When text received such attack, the relationship between watermark existence and converted text is shown as Fig 4.

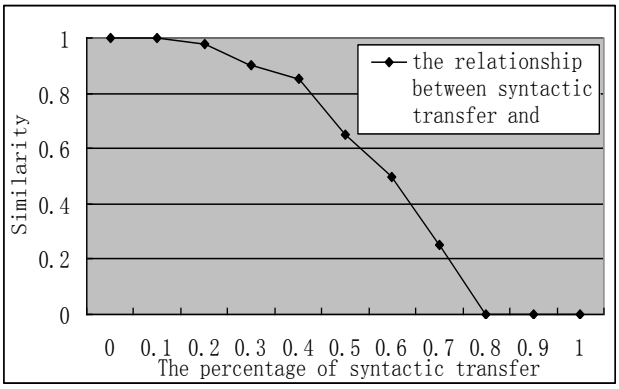

Figure 5. Watermark detection rate on syntactic transfer attack

From the fig 5 we can see that to the whole text, when sentences of the text has been syntactic transformed is equal or below 20%, the watermark existence rate can reach 100%, when the transform rate in the text equal to 50%, this algorithm can still maintain 60% of the watermark extraction rate. However, with the transform rate beyond 50%, the text will gradually lose its value. Based on the description of above, the algorithm has good robustness under the syntactic transformed attack.

C. Anti-attack Test

When use synonyms to replace part of the text words, the experimental results are shown below, synonyms used in the experiment are generated by the natural language processing technology. The result of the experiment is shown as Fig 6.

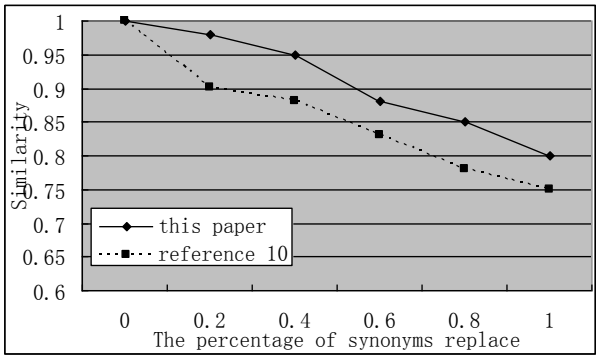

Figure 6. Watermark detection rate comparison on the synonyms replace attack

It can be seen from Fig 6, when synonym replace rate beyond 20%, this algorithm is superior to the algorithm of [10], especially, when the replacement rate exceeds 60%, the watermark existing probability is over 85%, it can still extract the zero-watermark effectively. According to this, under the synonym replace attack, our watermark detection algorithm efficiency has been significantly improved when compared with reference [10].

D. Robust Test

Robustness is an important index of watermark, the strength of the robustness directly influence the performance and the application of the watermarking algorithm, we use the text modification as an example for robustness analysis.

1) Modification Attack

In this experiment, modification divided into two types: insertion and deletion, we used 3 samples of different size text from data set designed in [15] to perform our experiments, these samples have been collected from Center for Chinese Linguistics PKU.

Insertion and deletion of data was performed at multiple randomly selected locations in text. Text category divided into 4 sizes: Small Size Text(SST), Medium Size Text(MST), Large Size Text(LST) and Very Large Size Text(VLST) respectively. Each sample was attacked with different insertion and deletion volume.

We evaluated the accuracy of extracted watermark with both localized and dispersed tampering attacks means 5%, 10%, 20% and 50% insertion attack, deletion attack and re-arranging of words on each of text samples. The average accuracy of extracted watermark under these tampering attacks is shown in table 1.

TABLE I.
THE ACCURACY OF WATERMARK
WITH INSERTION, DELETION RE-ARRANGING OF WORDS ATTACK

Text Category	Original Text	Watermark Accuracy under 3 ATTACKS			ATTACKS Volume
		Insertion	Deletion	re-arranging of words	
SST	283	98.3%	97.2%	96.7%	5%
		96.5%	95.8%	94.2%	10%
		88.6%	89.3%	84.2%	20%
		78.2%	76.9%	75.4%	50%
MST	549	94.8%	96.5%	95.4%	5%
		95.2%	94.8%	93.6%	10%
		88.7%	85.3%	84.2%	20%
		75.6%	72.2%	71.3%	50%
LST	1089	98.1%	97.8%	95.4%	5%
		95.8%	96.2%	91.2%	10%
		85.3%	82.4%	81.6%	20%
		77.2%	70.3%	75.4%	50%
VLST	4216	97.6%	93.4%	89.4%	5%
		83.2%	90.4%	89.6%	10%
		78.1%	82.4%	79.3%	20%
		76.2%	75.8%	75.1%	50%

As is shown in table 1, we can see that the accuracy of extracted watermark is always greater than 70%, and Small Size Text(SST) has the best performance with small attack volumes. For MST, it can be observed that watermark sensitive toward insertion attack.

In case of LST sample, the accuracy of extracted watermark is maximum in deletion attack and minimum under re-arranging attack.

Similarly, it can be seen from the table 1 that the VLST is more resilient towards different towards 3 different attacks since the accuracy is always above 75%.

In general, from the table 1 we can clearly see that our algorithm make a good performance under these modification attack.

2) Comparative Test

In order to further proof the algorithm presented in this paper has the good robustness, we carried out the comparative trial with reference [10]. The experiment use word document named "The life of blue sky". The total word is 1382. Comparison diagram of the experimental results is shown in Fig 7.

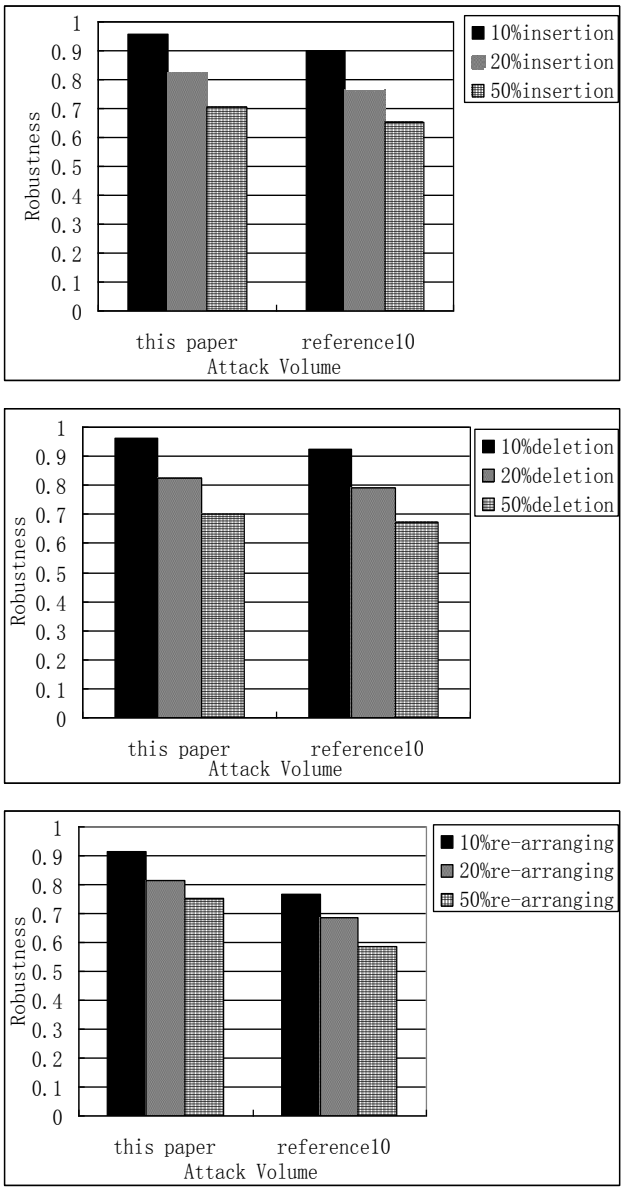

Figure 7. Comparision of robustness

It can be seen from Fig 7, robustness factor of watermark decreased with the text changes increasing. With different insertion, deletion and re-arranging of words attack volumes, this performance of algorithm in this paper is superior to the algorithm of [10]. Especially under re-arranging of words attack my algorithm has a significant advantage over the reference [10]. In practice, if a text has been modified more than 50% of the content, the text gradually lost its protective value. Therefore, the above two robust experiments show that this algorithm in robustness has certain advantages.

VI. CONCLUSION

In this paper, we design a text zero-watermark scenario by utilizing three-dimensional space model. Our algorithm combines general principles and methods of the traditional zero-watermark with the syntactic and semantic feature of text, and propose the usage of abstract-set in watermark

construction. Finally, the experiment evaluate and analyze the performance of the proposed algorithms under three different attacks with increasing attack volume. In comparison with other existing algorithms, our method has the better robustness and anti- aggressiveness.

In the further work, we can improve the final similarity calculation algorithm to make it more effective and our result will be improved at the same time. Moreover, our next step is to search a better method of extracting the semantic feature of text, so as to produce a more reasonable and general method to construct the abstract-set.

REFERENCES.

[1] J. T. Brassil, S. Low, N. F. Maxemchuk, and L. O'Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying," IEEE Journal on Selected Areas in Communications, vol 13, no. 8, October 1995, pp. 1495-1504.

[2] J. T. Brassil, S. Low, N. F. Maxemchuk, "Copyright protection for the election distribution of text document". Proceeding of IEEE, vol. 87, no. 7, July 1999, pp. 1181-1196

[3] N. F. Maxemchuk, S. H.Low. "Performance Comparision of Two Text Marking Methods," IEEE Journal of Selected Areas in Communication(JSAC), May 1998. vol. 16 no. 4 1998. pp. 561-572.

[4] N. F. Maxemchuk, "Electronic Document Distribution," AT&T Technical Journal, September 1994, pp. 73-80.6.

[5] S. H.Low, N. F. Maxemchuk, and A. M. Lapone, "Document Identification for Copyright Protection Using Centroid Detection," IEEE Transactions on Communications, Mar. 1998, vol. a5, no. 3, pp. 372-381

[6] M. J. Atallah, C. McDonough, S. Nirenburg, and V. Rakin, "Natural Language Pricessing for Inmation Assurance and Security: An Overview and Implementations", Proceedings 9th ACM/SIGSAC New Security Paradigms Workshop, September, 2000, Cork, Ireland, pp. 51-65.

[7] M. J. Atallah, V. Raskin, M. C. Crogan, C. F. Hempelmann, F. Kerschbaum, D. Mohamed, and SNaik, "Natural language watermarking: Design, analysis, and proof-of-concept implementation", Proceedings of the Fourth Information HidingWorkshop, vol. LNCS 2137, 25-27 April 2001, Pittsburgh, PA.

[8] P. Lu et al. "An optimized natural language watermarking algorithm based on TMR", on proceedings of 9th International Conference for Young Computer Scientists, 2009.

[9] Meng Yingjie, Si Lei, ShiYao, "A Chinese Texts Zero-Watermarking Algorithm Based on Vector Graphics". Journal of Computer Research and Development, 2009, Vol. 46(supp).22-26.

[10] Meng Yingjie, Guo Tao, Guo Zhihua, Gao Liming. "Chinese Text Zero-watermrk Based on Sentence's Entropy". ICMT2010: proceedings of the International Conference on Multimedia Technology, IEEE Computer Society, 864-867.

[11] Yanling Li, Jing Yuan; Xia Ye, "Method for Feature word weight caculating," on proceeding of Intelligent Computing and Intelligent Systems, ICIS 2009, PP:309-312.

[12] Kai-Ying Liu, Jia-Heng Zheng, "Research of automatic Chinese word Segmentation," on proceeding of Machine Learing and Cybernetics, ICMLC 2002, vol2, PP:805-809.

[13] Guoping Wang. An Efficient Implementation of SHA-1 Hash Function. In IEEE International.Conference on Electro/information Technology, 2006, PP:575–579.

[14] Sun X M, Luo G, Huang H J. Component-based digitial watermarking of Chinese texts[C]. Shanghai: Proc of the Third International Conference on Information Security, 2004: 76-81

[15] Z. Jalil and A. M. Mirza, "A Novel Text Watermarking Algorithm Based on Double Lettle", International Journal of Computer Mathematics.

Yingjie Meng born in 1964, Shanxi, China. He is currently a associate professor in LanZhou University. So far, he was published more than 20 papers in international conference proceedings and journals. His main research interests include data security, information processing.

Liming Gao born in 1986, Gansu, China. She is currently pursuing the M.S degree in LanZhou university. Her main research interests include data security, information processing.

Mingwen Liu born in 1987, Gansu, China. He is currently pursuing the M.S degree in LanZhou university. His main research interests include data security.

Lixin Bai born in 1982, Henan, China. She is currently pursuing the M.S degree in LanZhou university. Her main research interests include data security.