Word Length Based Zero-Watermarking Algorithm for Tamper Detection in Text Documents

Zunera Jalil, Anwar M. Mirza, and Hajira Jabeen
FAST National University of Computer and Emerging Sciences,
A.K. Barohi Road, H-11/4, Islamabad, Pakistan
E-mail: {zunera.jalil, anwar.m.mirza, hajira.jabeen}@nu.edu.pk

Abstract- **Copyright protection and authentication of digital content has become a major concern in the current digital era. Plain text is the widely used means of information exchange on the Internet and it is essential to verify the authenticity of information in any form of communication. There are very limited techniques available for plain text watermarking, authentication, and tamper detection. This paper presents a novel zero-watermarking algorithm for tamper detection in plain text documents. The algorithm generates a watermark based on the text contents which can be extracted later using extraction algorithm to identify the status of tampering in the text document. Experimental results demonstrate the effectiveness of the algorithm against random tampering attacks. Watermark pattern matching and watermark distortion rate are used as evaluation parameters on multiple text samples of varying length.**

Keywords- watermarking; tamper detection; authentication; security; algorithm

I. INTRODUCTION

The copyright protection and authentication of digital content has become more important with the increasing use of Internet, e-commerce, and other communication technologies effectively. Besides, making it easier to access information in a short time span, it has become difficult to protect copyrights of digital content and to prove the authenticity of the obtained information. Digital contents mostly comprises of text, image, audio and video. Authentication and copyright protection of digital image, audio, and video has been given due thought by the researchers in past. However, authentication, tamper detection, and copyright protection of plain text have been ignored. Most of the digital contents like websites, e-books, articles, news, chats, SMS, are in the form of plain text.

The threats of illegal copying, tampering, imitation, plagiarism, forgery, and other forms of possible disruption need to be exclusively addressed for the plain text. Digital watermarking provides a solution to authenticate and to protect digital contents. Digital watermarking methods are used to identify the original copyright owner (s) of the contents which can be an image, a plain text, an audio, a video or a combination of all.

A digital watermark is visible or invisible (preferably the later) identification code that is permanently embedded in the data. It means that unlike conventional cryptographic techniques, it remains present within the data even after the decryption process [1].

A text is the easiest mode of communication and information exchange, brings many challenges when it comes to copyright protection and authentication. All changes to the text must preserve the value, utility, meaning and grammaticality of the text. Short documents are harder to protect and authenticate since a simple analysis would easily reveal the watermark.

In image, audio, and video watermarking the limitations of Human Visual and/or Human Auditory System and inherent redundancies are exploited for watermark embedding. It is difficult to find such limitations and redundancy in plain text, since text is sensitive to any modification required for watermark embedding.

Text is easier to copy, reproduce and tamper as compared with images, audio and video. Text being a specialized medium requires specialized copyright protection and authentication solutions. Traditional watermarking algorithms modify the contents of the digital medium to be protected by embedding a watermark. This traditional watermarking approach is not practical for plain text. A specialized watermarking approach such as zero-watermarking would do the needful for plain text. In this paper, we propose a novel zero- watermarking algorithm which utilizes the contents of text itself for its authentication. A zero-watermarking algorithm does not change the characters of original data, but utilize the characters of original data to construct original watermark information [2-3].

The paper is organized as follows: Section 2 provides an overview of the previous work done on text watermarking. The proposed embedding and extraction algorithm are described in detail in section 3. Section 4 presents the experimental results for the tampering (insertion, deletion and re-ordering) attacks. Performance of the proposed algorithm is evaluated on multiple text samples. The last section concludes the paper along with directions for future work.

II. PREVIOUS WORK

Text watermarking for authentication of text documents is an important area of research; however, the work done in this domain in past is very inadequate. The work on text watermarking initially started in 1991. A number of text watermarking techniques have been proposed since then. These include text watermarking using text images, synonym based, pre-supposition based, syntactic tree based, noun-verb based, word or sentence based, acronym based, typo error based methods and many others.

978-1-4244-6349-7/10/$26.00 ©2010 IEEE

Broadly speaking, the previous work on digital text watermarking can be classified in the following categories; an image based approach, a syntactic approach, and the semantic approach. Description of each category and the work done accordingly is as follows:

A. *Image-Based Approach*

In image based approach towards text watermarking, the image of text is taken as a source for watermark embedding. Brassil, et al. were the first to propose few text watermarking methods utilizing text image[4]-[5]. Later Maxemchuk, et al. [6]-[8] analyzed the performance of these methods. Low, et al. [9]-[10] further analyzed the efficiency of these methods. The first method was the line-shift algorithm which moves a line upward or downward (left or right) based on watermark bit values. The word-shift algorithm used the inter-word spaces to embed the watermark. The last method was the feature coding algorithm in which specific text features are tampered to encode watermark bits in the text.

Huang and Yan [11] proposed an algorithm based on an average inter-word distance in each line. The distances are adjusted according to the sine-wave of a specific phase and frequency. The feature and the pixel level algorithms were also developed which mark the documents by modifying the stroke features such as width or serif [12].

B. *Syntactic Approach*

In this approach, the syntactic structure of the text is used to embed watermark. Mikhail J. Atallah, et al. proposed the first natural language watermarking scheme by using syntactic structure of text [13]-[14] where the syntactic tree is built and transformations are applied to it in order to embed the watermark keeping all the properties of text intact. The NLP techniques are used to analyze the syntactic and the semantic structure of text while performing any transformations to embed the watermark bits.

Hassan et al. performed morpho-syntactic alterations to the text to watermark it [15]. The text is first transformed into a syntactic tree diagram where text hierarchy and dependencies are analyzed to embed watermark bits. Hassan et al. provided an overview of available syntactic tools for text watermarking [16].

C. *Semantic Approach*

The semantic watermarking schemes focus on using the semantic contents of text to embed the watermark. Atallah et al. were the first to propose the semantic watermarking schemes in the year 2000 [17]-[19]. Later, the synonym substitution method was proposed, in which watermark is embedded by replacing certain words with their synonyms [20]. Xingming, et al. proposed noun-verb based technique for text watermarking [21] where nouns and verbs in a sentence are parsed using grammar parser and semantic networks. Later Mercan, et al. proposed an algorithm of the text watermarking by using typos, acronyms and abbreviation to embed the watermark [22]. Algorithms were developed to watermark the text using the presuppositions [23] by observing the discourse structure, meanings and representations. The text pruning and the grafting algorithms were also developed in the past. The algorithm based on text meaning representation (TMR) strings has also been proposed [24].

The above mentioned approaches and algorithms are not applicable to all types of text documents under random tampering attacks and are not designed specifically to solve tamper detection problem; hence we propose a zero-watermarking algorithm which incorporates the contents of text for its protection and tamper detection.

III. PROPOSED ALGORITHM

Traditional watermarking approaches for text documents aims at embedding additional information in text and this information can be used later for authentication, tamper detection or copyright protection. We propose a zero-watermarking approach in which the host text document is not altered to embed watermark, rather the characteristics of text are utilized to generate a watermark. The contents of text are utilized to generate a watermark. This watermark pattern is later matched using a pattern matching procedure with the pattern generated by tampered document to identify any tampering.

The watermark generation and extraction process is illustrated is fig. 1. Watermark is registered with the Certifying Authority (CA) with current date and time and is used in the extraction algorithm later to detect tampering in the text document.

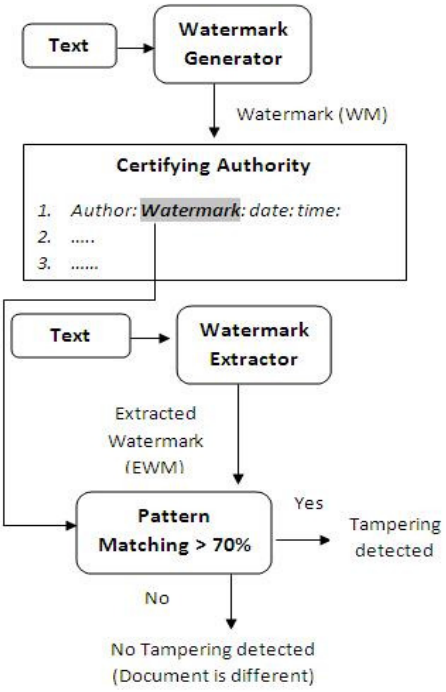

Figure 1. Overview of Watermark Generation and Extraction Processes

The proposed algorithm utilizes the contents of text to protect it. Generally, an attacker performs meaning preserving tampering in text and the intention is to make it look different. Tampering can be passivization, clefting, topicalization or re-phrasing. All these tampering do not

alter the nouns, adjectives or proverbs which usually contain more than 4 letters. The attacker shuffles the text (preserving meaning) and changes the placement of these nouns/adjectives but cannot avoid/skip them. All the words having more than four letters are identified and the initials are used to generate watermark patterns. These patterns are concatenated to construct a watermark. This process is illustrated in fig. 2, where a watermark is generated based on text contents.

Sentences	Pattern
Pakistan is a developing country.	PDA
Islamabad is the capital city of Pakistan.	ICP
Islamabad is located in the Pothohar Plateau in the north of the country.	ILPPNC
Islamabad is one of the greenest and most well-planned cities of South Asia.	IGWCS

Watermark = PDA.ICP.ILPPNC.IGWCS.

Figure 2. Watermark Generation

It is a zero-watermarking algorithm, since watermark is not actually embedded in the text itself, rather it is generated by using the characteristics of text. A typical watermarking process involves two stages: (1) embedding algorithm and (2) extraction algorithm. Watermark embedding is done by the original author and extraction done later by a Certifying Authority (CA) to detect tampering and/or prove ownership. This trusted Certifying Authority (CA) is necessary in this algorithm with whom, the original copyright owner/author of the text registers his/her watermark. Whenever the content/text ownership is in question or authentication need to be proved, this trusted third party acts as a neutral decision authority.

A. Embedding Algorithm

The algorithm which embeds the watermark in the text is called embedding algorithm. The watermark embedding algorithm requires original text file as input. A watermark is generated as output by this algorithm. This watermark is then registered with the certifying authority along with the original text document, author name, current date and time. The original text (T_O) be first obtained from the author and length of each word in each sentence of the text is analyzed. All words having more than 4 letters are identified and their initial letter is used to generate watermark patterns. All patterns are concatenated to generate the watermark. This watermark is then registered with the CA with current date and time.

The algorithm proceeds as follows:

```
1. Read text file T0.
2. NS = Total number of sentences in T0
3. for i=1 to NS, repeat step 4 to 9.
4. NW=Number of words in ith sentence
5. for j=1 to NW, repeat step 6 to 8
6. LW= Length of jth word
7. if ( LW > 4)
      W(i,j)= First letter of jth word
8. j = j+1
9. i = i+1
10. Output W (watermark)
```

B. Extraction Algorithm

The algorithm which extracts the watermark from the text is called the extraction algorithm. The proposed extraction algorithm takes the plain text as input. The text may be attacked or un-attacked. The watermark pattern is generated from the text by the extraction algorithm. Each sub pattern of this watermark is then compared with the corresponding original watermark sub pattern registered with the CA.

Each pattern of original watermark is compared with the primary and secondary patterns of the corresponding extracted watermark. Primary matching is performed first. If it retrieves the exact watermark, then document is said to be intact. In case primary matching is unsuccessful, secondary matching is performed which compares the extracted watermark pattern with all possible combinations of original watermark pattern. The process is shown is fig. 3.

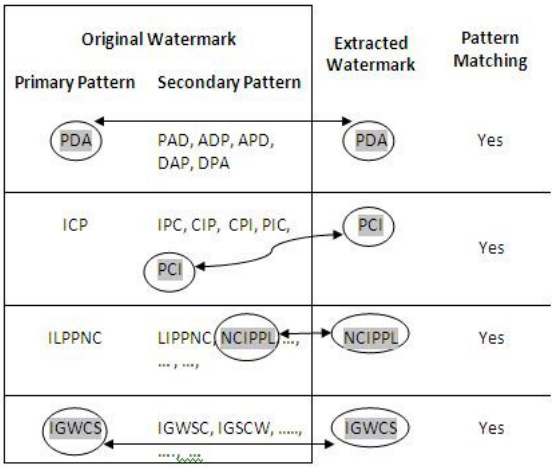

Figure 3. Watermark extraction process

The watermark will be accurately detected by this algorithm in the absence of attack on text, and text document will be called authentic text without tampering. Watermark will resist common sentence re-writing attacks. However, it will get destroyed with extensive tampering attacks. Tampering can be insertion, deletion, paraphrasing or re-ordering of words and sentences in text. The extraction algorithm is as follows:

```
1. Read text file T_A and OW
2. NS = Total number of sentences in T_A
3. for i=1 to NS, repeat step 4 to 9.
4. NW=Number of words in ith sentence
5. for j=1 to NW, repeat step 6 to 8
6. LW= Length of jth word
7. if ( LW > 4)
      EW(i,j)= First letter of jth word
8. j = j+1
9. I = i+1
10. if EW = OW (primary match)
      PMR(p) = 1
    else
      if EW = OW (secondary match)
      PMR(s)= No. of matched secondary
              patterns/TP
    else
      PMR= (NM(p)+NM(s))/TP

T_A=Attacked text; PMR= Pattern matching
rate; TP = Total patterns; OW = Original
watermark; EW = Extracted watermark;
NM = Number of matched patterns
```

IV. Experimental Results

We used 6 samples of variable size text from the data set designed in [25] for our experiments. These samples have been collected from Reuters' corpus, e-books, and web pages. Insertion and deletion of words and sentences was performed at multiple randomly selected locations in text. Table I show the sample number as in dataset [25], number of words (WC_o) in original text, the insertion and deletion volume, and the number of words (WC_A) in the text after attack.

TABLE I
ORIGINAL AND ATTACKED TEXT SAMPLES WITH INSERTION AND DELETION RATIOS

Sample No.	Original Text	Attack		Attacked Text
	WC_O	Insertion	Deletion	WC_A
1 : [SST2]	421	26%	25%	425
2 : [SST4]	179	44%	54%	161
3: [MST2]	559	49%	25%	696
4: [MST4]	2018	14%	12%	2048
5: [MST5]	469	57%	53%	491
6: [LST1]	7993	9%	6%	8259

Pattern Matching Rate (PMR) and Watermark Distortion Rate (WDR) are calculated as per the following formulas:

$$PMR = \frac{\text{Number of patterns correctly matched}}{\text{Number of watermark patterns}}$$

$$WDR = 1 - PMR$$

The values of PMR ranges between 0 (the lowest) and 1(the highest) with desirable value close to 1. The values of WDR also ranges between 0 (the highest) and 1(the lowest) with value close to 0 as desirable value. PMR threshold was set to 0.7 to detect tampered document. PMR values less than 50, less than 70 and greater than 70 represents less, moderate and high level of tampering. Pattern matching rate and watermark distortion rate are shown in table II.

TABLE II
ACCURACY OF EXTRACTED WATERMARK WITH KEYWORD 'AND'

Sample No.	PMR	WDR	Tamper Detection
1 : [SST2]	0.5671	0.4329	High
2 : [SST4]	0.8634	0.1366	Less
3: [MST2]	0.7941	0.2059	Less
4: [MST4]	0.8335	0.1665	Less
5: [MST5]	0.6332	0.3668	Moderate
6: [LST1]	0.8548	0.1452	Yes

It can be observed in table II that tampering with text is always detected, whether it is low, moderate or high. Watermark distortion rate greater than 0 indicate that watermark has been distorted as a result of tampering.

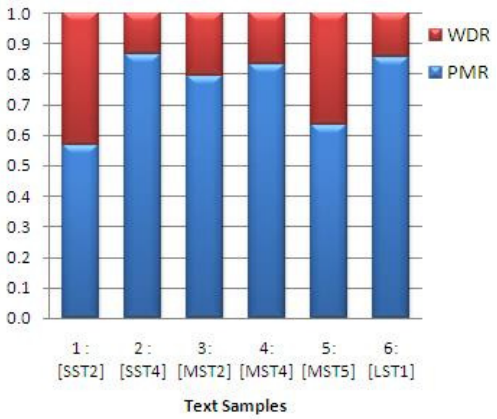

Figure 4. Pattern matching and watermark distortion rate for all text samples

Figure 4 shows the watermark pattern matching (PMR) and watermark distortion rate (WDR) for all text samples. It can be clearly observed that pattern matching rate clearly indicate low, moderate and high state of tampering where tampering with text always gets detected. Text is sensitive to meaning preserving modifications made by the attacker to make it look different or to destroy the writing style of original author. Positive distortion rate indicates that the text has been tampered and is not authentic. This proves that the accuracy of watermark gets affected even with minor tampering and the evident watermark fragility proves that text has been attacked.

V. Conclusion

The existing text watermarking solutions for text authentication are not applicable under random tampering attacks and on all types of text. With the small volume of attack, it becomes impossible to identify the existence of watermark and to prove the authenticity of information. We have developed a zero-text watermarking algorithm, which utilizes the contents of text to generate a watermark and this watermark is later extracted to prove the authenticity of text

document. We evaluated the performance of the algorithm for random tampering attack in dispersed form on 6 variable size text samples. Results show that our algorithm always detects tampering even when the tampering volume is low. This work can further be extended to include the multiple occurrences of letters in a single word; for watermark embedding.

ACKNOWLEDGMENT

One of the authors, Ms. Zunera Jalil, 041-101673-Cu-014 would like to acknowledge the Higher Education Commission of Pakistan for providing the funding and resources to complete this work under Indigenous Fellowship Program.

REFERENCES

[1] A. Khan, A. M. Mirza and A. Majid, "Optimizing Perceptual Shaping of a Digital Watermark Using Genetic Programming", Iranian Journal of Electrical and Computer Engineering, vol. 3, pp. 144-150, 2004.

[2] A. Li, B. Lin, Y. Chen, "Study on Copyright Authentication of GIS vector data based on Zero-watermarking", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. VII. Part B4, pp.1783-1786, 2008.

[3] X. Zhou, W. Zhao, Z. Weidong, L. Pan, "Security Theory and Attack analysis for Text Watermarking", 2009 International Conference on E-Business and Information System Security, EBISS 2009.

[4] 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.

[5] J. T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright Protection for the Electronic Distribution of Text Documents," Proceedings of the IEEE, vol. 87, no. 7, July 1999, pp.1181-1196.

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

[7] N. F. Maxemchuk, "Electronic Document Distribution," AT&T Technical Journal, September 1994, pp. 73-80. 6.

[8] N. F. Maxemchuk and S. Low, "Marking Text Documents," IEEE International Conference on Image Processing, Washington, DC, Oct. 26-29, 1997, pp. 13-16.

[9] 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. 46, no.3, pp 372-381.

[10] S. H. Low and N. F. Maxemchuk, "Capacity of Text Marking Channel," IEEE Signal Processing Letters, vol. 7, no. 12 , Dec. 2000, pp. 345 -347.

[11] D. Huang and H. Yan, "Interword distance changes represented by sine waves for watermarking text images," IEEE Trans. Circuits and Systems for Video Technology, Vol.11, No.12, pp.1237-1245, Dec 2001.

[12] T. Amano and D. Misaki, "A feature calibration method for watermarking of document images," Proceedings of ICDAR, pp.91-94, 1999.

[13] M. J. Atallah, C. McDonough, S. Nirenburg, and V. Raskin, "Natural Language Processing for Information Assurance and Security: An Overview and Implementations", Proceedings 9th ACM/SIGSAC New Security Paradigms Workshop, September, 2000, Cork, Ireland, pp. 51–65.

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

[15] H. M. Meral et al., "Natural language watermarking via morphosyntactic alterations", Computer Speech and Language, 23, 107-125, 2009.

[16] H. M. Meral, et al., "Syntactic Tools for Text Watermarking", 19th SPIE Electronic Imaging Conf. 6505: Security, Steganography, and Watermarking of Multimedia Contents, Jan. 2007, San Jose, USA.

[17] M. Atallah, C. McDonough, S. Nirenburg, and V. Raskin, "Natural Language Processing for Information Assurance and Security: An Overview and Implementations," Proceedings 9th ACM/SIGSAC New Security Paradigms Workshop, September, 2000, Cork, Ireland, pp. 51–65.

[18] M. Atallah, V. Raskin, C. F. Hempelmann, M. Karahan, R. Sion, U. Topkara, and K. E. Triezenberg, "Natural Language Watermarking and Tamperproofing", Fifth Information Hiding Workshop, vol. LNCS, 2578, October, 2002, Noordwijkerhout, The Netherlands, Springer-Verlag.

[19] M. Topkara, C. M. Taskiran, and E. Delp, "Natural language watermarking," Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VII, 2005.

[20] U. Topkara, M. Topkara, M. J. Atallah, "The Hiding Virtues of Ambiguity: Quantifiably Resilient Watermarking of Natural Language Text through Synonym Substitutions". In Proceedings of ACM Multimedia and Security Conference, Geneva, 2006.

[21] X. Sun, A. J. Asiimwe, "Noun-Verb Based Technique of Text Watermarking Using Recursive Decent Semantic Net Parsers", Lecture Notes in Computer Science (LNCS) 3612: 958-961, Springer Press, August 2005.

[22] M. Topkara, U. Topraka, M. J. Atallah, "Information Hiding through Errors: A Confusing Approach". In: Delp III, E.J., Wong, P.W. (Eds.), Security, Steganography, and watermarking of Multimedia Contents IX. Proceedings of SPIE-IS&T Electronic Imaging SPIE 6505. pp. 65050V-1–65050V-12, 2007.

[23] B. Macq and O. Vybornova, "A method of text watermarking using presuppositions," in Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents, January 2007.

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

[25] Z. Jalil and A.M. Mirza, "A Novel Text Watermarking Algorithm Based on Double Letters", International Journal of Computer Mathematics. (Submitted)