A Text Watermarking Algorithm Developed Using Natural Language Processing

Khandokar Nafis Jaman[1], Zannatun Nayem[1], Bristi Rani Roy[1], Nayreet Islam[1], Faisal R. Badal[2], Subrata K. Sarker[2]

[1]Department of Computer Science and Engineering, Bangladesh Army University of Engineering and Technology, Nator, Bangladesh

[2]Department of Mechatronics Engineering, Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh

Abstract:-*Digital watermarking is a new technique for protecting copyright digital product. In modern times digital watermarking is mainly focused on audio, image, and video. This paper plans and accomplishes a digital text watermarking algorithm. In digital text watermarking, several techniques are accomplished for Chinese, English, Arabic and Turkish language text by using several methods. This paper covers a new digital watermarking technique which applicable for text documents in the English language. An algorithm has been proposed by this paper which is founded on grammatical words in English and a proper encryption procedure. The grammatical rules have been focused by us like article, preposition, modal verbs and conjunction to produce encrypted message by using a watermark. The produced technique is used for verifying the several websites.*

Keywords: *Text watermarking; Natural language processing; RSA; Encryption; Author's authenticity*

1. INTRODUCTION

We are living in the modern era. Comparatively, people are mostly depended on internet based things. Meanwhile, the number of users of the internet has reached 1,574 Million. Eventually, more digital text information is available than hard copies. Alongside the development of technology, people are becoming more comfortable with digital text information than hard copies. Because of handling, sharing, gathering or getting, digital text information is more easy and efficient. Nevertheless, it is really important to provide enough security to this digital text information. In other words, Internet technology's development and digital media's supremacy not only fetch great amenities for us but also malicious assault illegal piracy which caused copyright palavers problems [1]. Besides these digital text information can be illegally copied or have some threat like important data theft, authentication problems, and forgery. Basically, there are various solutions or many effective ways to overcome these threats. Confidentiality, authenticity, integrity can be used to overcome these threats [2]. But providing protection against copyright is a great challenge, maybe it is tough or nearly impossible to prevent copyright totally. Yet it is possible to protect copyright or detect it by using various methods. Digital watermarking method is used to provide protection against digital text information copyright. In this method secret information of data embedded into original digital text then compare it and find out owner attribution of data and tracking infraction [3]. However different kind of digital watermarking method has been developed recently. Some method focuses on or finds out attribution of the owner using semantic or syntactic analysis, text format, line space, characters or words font [2].Based on its type of carrier digitalwatermarking can be divided into image, video, audio, text and other media [4]. The main advantages of digital watermarking are, it is efficient, robust, effective and easy to use. At the same time, it provides protection against different threats and helps to detect the copyright part and also the real owner of that part of the text. In a nutshell digital watermarking is very effective and tremendous obtainment to provide protection against threats and detection of copyright.

In this paper a novel algorithm has suggested for digital watermarking, which is mainly work with Parts of Speech tags including modal verbs, prepositions, conjunctions and articles. Here RSA algorithm is used as encryption technique. This suggested algorithm is very efficient and useful to detect the copyright of document. This algorithm also able to show the real author name form which, text document is copied with the highest accuracy. And also provide necessary protection against various threats.

2. LITERATURE REVIEW

Makarand L. Mali, Nitin N. Patil and J. B. Patil [2] proposed a watermarking algorithm based on English grammatical words and used encryption technique. In this approach to generate encrypted message grammatical rules like modal verbs, pronouns and conjunctions have been used.

Chen Li and You Fucheng [4] implemented a novel algorithm which is text digital watermarking algorithm based on the word document, by modifying the font of the letters in the word document. Finally, through a dilution of the test, it explores and shows the accuracy and efficiency of the algorithm.

He Lu et al. [5] in text watermarking implemented an unprecedented method by using inter-character spaces and also background texture to represent noise if and only if OCR attack is recognized.

Yingli Zhang et al. [6] has proposed a very efficient method of watermarking based on the word document for controlling the dissemination and conserving copyright, for both Chinese and English language. In this approach, each object of word document contains information of author and legal user after performing encryption technique, grouping and packing into the message.

C. Culnaneet. al. [7] has suggested a watermarking method for formatted text documents. In this method, word spaces are used and take the documents as one long line for watermarking. Nevertheless, the author proposed a unique method of threshold and thresholding buffering.

Xianghe Jing, HuapingFei, YuHao and Zhijun Li [8] proposed a novel text encryption method based on natural languageprocessing (NLP). Three linguistic transformation Synonym substitution, Syntactic transformations, Semantic transformations are introduced and new encryption technique is provided.

Daojing Li and Bo Zhang [9] have implemented a Dual Watermarking method founded on ambit Cryptography (DWTC) for Web information to solve the cruxes of robustness and invisibleness. This job founded on ambit cryptography, watermarking method can improve the toughness.

MercanTopkara et al. [10] has suggested a natural language watermarking by applying sentence composition to apply a watermark The text phrase compositions such as characters, words and lines were modified to apply the necessary information. The authors provide an audit of governing status of the efficiency in natural language watermarking, tools and techniques for text processing.

3. DIGITAL WATERMARKING TECHNIQUE

A digital watermarking is a kind of signal or string embedded in a noise-tolerant signal which is usually identified the proprietary right of the copyright of these signal. It is generated by applying the different embedded algorithm. Various categories of algorithm or technique are derived or generated by the researchers.

3.1 Document Structure

3.1.1 Space Coding

Two types of space coding are typically used. Line space and word space. Line space is worked with a space between two adjacent rows of a paragraph. Watermark is embedded by tactfully changing the space of the adjacent line. Though having strong robustness and difficult to track watermark, the capacity of the watermark is very small and difficult for visualization. Another space coding technique is word space coding. Word coding work with horizontal movement of the word. To embed watermark, word of the same row are shifting left or right. Invisible coding is one of the approaches to space coding. Watermark information is attached at the line break. But it is difficult for visualization whether it is tab or space at the end of the line.

3.1.2 Feature Coding

Changing particular features like font-family, indent, color, text-style, font-style are the main scheme of feature coding. As a consequence of the various type of information, information capacity of the watermark are usually vast in comparison to another space coding.

3.2 Natural Language Processing

3.2.1 Synonym substitution

One of the simplest and most widely used embedded watermarking technique is synonym substitution. To keep unchanged the meaning of the sentence, synonym word is used to substitute the word.

3.2.2 Syntactic transformation

Syntactic transformation mainly works with the syntactic transformation of a sentence. Sometimes it does a little bit change of meaning of the sentence. Conversion of active to passive or vice versa, adding a formal subject, slicing a sentence, placing topic atthe beginning of the sentence is some approach of syntactictransformation.

3.2.3 Semantic transformation

In semantic transformation, data representation is changed from one model to another using semantic information. Different types of semantics technique are used. Words contained the same meaning are being pruned, replace a word with same meaningful word or phrase are few approaches of semantic transformations.

4. METHODOLOGY

As an increasing number of information distribution on the internet, illegitimate use of data use, transfer, copy or distribution become one of the most alarming issues for the authors and writers. To secure the original authorship and copyright, the necessity of digital text watermarking is raising upward. To protect authorship and copyright along with the original formation of data, a robust strong watermarking algorithm is needed.

In this paper, a strong and more robust algorithm is proposed for digital watermarking based on Natural Language Processing technique. The proposed algorithm work with Parts of Speech (POS) tags including modal verbs, prepositions,conjunctions and articles present in the text document.

To form the watermarks, at first, this method scarp a web page and sum up the number of total occurrences of modal verbs, prepositions, conjunctions and articles of the text document. Then we convert the number of occurrences into binary and concatenate this binary number with author's ID. This author's ID can be given by Certified Authority (CA) or the author can generate himself. Finally, RSA algorithm is applied to this combined string to form up an encrypted key. The length of this encrypted key is not certain. This technique can be secure the authorship and copyright of the original author. To extract the watermark from the actual text, the opposite way of this procedure can be applied.

4.1 Mathematical Formula

Let

$n(p)$ = Number of total occurrences of preposition in the text document.

$n(c)$ = Number of total occurrences of conjunction in the text.

$n(mv)$ = Number of total occurrences of modal verb in the text.

$n(a)$ = Number of total occurrences of article in the text.

$AuthID$ = Author's ID

Step-1

$$key = \left\{ \sum_{length=1}^{n} n(p) + n(c) + n(mv) + n(a) + AuthID \right\}$$

Step-2

$$key = (key)_{binary}$$

RSA algorithm is applied to this combined key for final encryption to generate the watermark. This algorithm is implemented in python language.

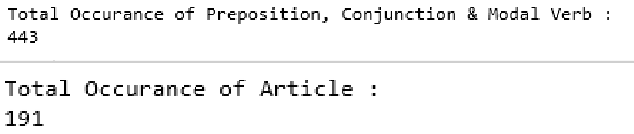
5. RESULTS

In this above experiment, we use three web page's content to generate watermark by applying our proposed algorithm. The result is given below:

5.1 Case Study 1

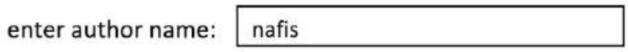
First web page: *http://walthowe.com/navnet/history.html*

This page contains a text document about the internet. We scrap this web page to get contents and apply our proposed algorithm to this content. A unique key is generated by applying our method. Figure 1, Figure 2 shows the process.



```
Total Occurance of Preposition, Conjunction & Modal Verb :
443
Total Occurance of Article :
191
```

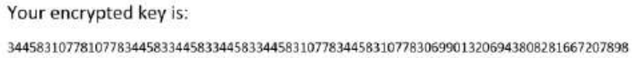
Fig -1: Number of watermark



enter author name: nafis

Fig -2: Author's ID

Figure 3 shows the encrypted key generated by applyingour algorithm.

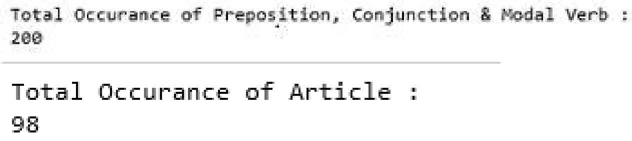


```
Your encrypted key is:
3445831077810778344583344583344583344583107783445831077830699013206943808281667207898
```

Fig -3: Generated key

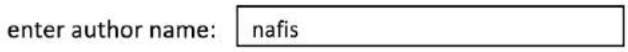
5.2 Case Study 2

We use another web page for our experiment. Second web page: *https://nytcrossword.com/tag/it-may-allow-a-text-document-to-be-displayed-on-a-web-page-crossword-clue*



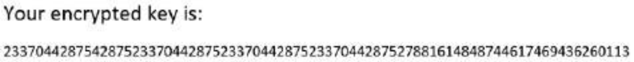
```
Total Occurance of Preposition, Conjunction & Modal Verb :
200
Total Occurance of Article :
98
```

Fig **-4**: Number of watermark



enter author name: nafis

Fig -5: Author's Id



```
Your encrypted key is:
233704428754287523370442875233704428752337044287527881614848744617469436260113
```

Fig -6: Generated Key

Figure 5, Figure 6 show the process of the algorithm. Figure 6 shows the encrypted key generated by applying our algorithm.

6. CONCLUSION

In this paper, we have implemented a newish approach of digital watermarking for text document by some natural language watermark and grammatical rules. We have applied this technique in several web pages and get a better result of generating the watermark. The resultant encrypted watermarkof this algorithm is found to be much more secured and strong that can help the actual author to protect their authorship and copyright. Convert amounts of the watermark to binary and applied RSA algorithm made this key more robust. For every English text document, it is found very effective and compatible. Our proposed algorithm is implemented with python language because of its compatibility to work with Natural Language Processing.In this paper, we implemented this algorithm for the English language. In the future, we will try to make consistent with other languages (especially Latin) with this algorithm.

REFERENCES

[1] Wang Zhigang, *Rearch of Watermarking algorithm for WORD Document,*China Science and Technology Information, Mar.2010, pp.114.

[2] Makarand L. Mali, Nitin N. Patil, J. B. Patil, *Implementation of Text Watermarking Technique Using Natural Language Watermarks,* 2013 International Conference on Communication Systems and Network Technologies.

[3] Chen Li, You Fucheng, *The Study on Digital Watermarking Based on Word document,* 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC) Dec 20-22, 2013, Shenyang, China.

[4] Chen Qing, Zhou Limin, *The Research of Digital Watermarking Algorithm Based on WORD Document Image Processing,* 2010, pp. 271–350.

[5] He Lu, Fang Ding Yi, Gui Xiao Lin, Chen Xiao Jiang, XuXinBai, Liu Jin'An, *A New Chinese Text Digital Watermarking for Copyright Protecting Word Document,* 2009 International Conference on Communications and Mobile Computing, 978-0-7695- 3501-2/09

[6] Yingli Zhang, Huaiqing Qin, *A Novel Robust Text Watermarking For Word Document 3rd International Congress on Image and Signal Processing,*Vol. 1, pp. 38-42, October 2010.

[7] C. Culnane, H. Treharne, and A.T.S. Ho, *Improving Multi-Set Formatted Binary Text Watermarking Using Continuous Line Embedding,* in Proceedings of IEEE International Conference on Innovative Computing, Information and Control (ICICIC-07), Kumamoto, Japan, pp. 287-29, 2007.

[8] Xianghe Jing, Yu Hao, HuapingFei, Zhijun Li,*Text Encryption Algorithm Based on Natural Language Processing,* 2012 Fourth InternationalConference on Multimedia Information Networking and Security

[9] Daojing Li and Bo Zhang, *DWTC: A Dual Watermarking Scheme Based on Threshold Cryptography for Web Document,* 2010 International Conference on Computer Application and System Modeling (ICCASM 2010)

[10] M. Topkara, *New Designs for Improving the Efficiency and Resilience of Natural Language Watermarking,* PhD Thesis, Purdue University, WestLafayette, Indiana, 2007