A Method of Generating and Tracing Unmarked Text Watermarking Based on Thesaurus

Bohan, Wang*
School of Computer Science and Technology, Huazhong University of Science and Technology
bohan_wang@hust.edu.cn

Yifei, Yu
School of Computer Science and Technology, Huazhong University of Science and Technology
u202015329@hust.edu.cn

Hefei, Ling
Huazhong University of Science and Technology
lhefei@hust.edu.cn

ABSTRACT

Text watermarking technology has been applied broadly to trace the source of data leakage and achieve copyrights protection. However, modifying text format or text structure, the two essential methods of watermarking technology, lack resistance in the application of OCR and other identification technologies, for it removes both the format and the structure of the text. Therefore, this paper proposes an OCR-resistant text watermarking technique independent of the structure and the format in Chinese, and explores the tracing watermarking of such unmarked text. Firstly, binary watermark sets are generated based on parity check. Then, the dictionary set of N groups is established on synonym pairs denoted by 0 and 1 respectively. Lastly, the text to be encoded is sequentially iterated over and the keywords in the text are replaced by the corresponding synonyms in the thesaurus according to the binary watermark bit. The method presented in this paper can effectively resist the invalid watermark caused by the modification of the format and the structure. Even after the text is scanned or reformatted by OCR, the watermark still exists. At the same time, invisible tracing watermarking secures watermark protection and text tracing because parity check can enhance the error detection capability and separated word database used to generate index can improve the operational performance. Experiments show that this method has good invisibility and high resistance to various text attacks, including format transformation, structure transformation, and retyping the text.

CCS CONCEPTS

• **Security and privacy**; • **Software and application security**; • **Domain-specific security and privacy architectures**;

KEYWORDS

Text watermarking, steganographic tracing, synonym dictionary, parity check

*Corresponding author

HP3C 2023, June 17–19, 2023, Jinan, China

ACM ISBN 978-1-4503-9988-3/23/06…$15.00
https://doi.org/10.1145/3606043.3606065

ACM Reference Format:
Bohan, Wang, Yifei, Yu, and Hefei, Ling. 2023. A Method of Generating and Tracing Unmarked Text Watermarking Based on Thesaurus. In *2023 7th International Conference on High Performance Compilation, Computing and Communications (HP3C) (HP3C 2023), June 17–19, 2023, Jinan, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3606043.3606065

1 INTRODUCTION

Digital watermarking technology has been applied broadly in marking and tracing data in copyrights protection. The key of this technology lies in its invisibility and robustness of watermark, namely, it is invisible to the naked eye but can be detected by the algorithm. Therefore, attacks such as cutting, splicing and editing can be resisted to a certain extent.

Electronic text carries such an abundant information that embedding of digital watermarking for text data is extremely important. Currently, there are 4 text watermarking technologies, including watermarking algorithm based on document format, watermarking algorithm based on document structure, watermarking algorithm based on traditional binary image processing, and watermarking algorithm based on natural language processing [Zhao et al. 2020]. The watermarking algorithm based on document format is encoded and embedded on the basis of the characteristics of different formats [Brassil et al. 1999; Ding and Hong 2001; Hui ; Li and Zhu 2008; Luo et al. 2008; Xu 2013]. The watermarking algorithm based on document structure is to encode and embedded on the basis of the characteristics of different structures [Gu and Feng 2015; Kang 2008; Liu et al. 2006; Xu 2013]. They above two algorithms are simple and suitable for text with typesetting function, such as WORD and PDF, but they are strongly dependent on the text format and lack of robustness. The watermarking algorithm based on traditional binary image processing is to convert the text into a binary image document, and then the information is hidden between different digital blocks [Wu and Liu 2004; Zhang 2007; Zhao et al. 2008]. However, the algorithm subtly processes the pixels of individual characters and generate visually similar new characters, which might lead to the ignorance of pixels in image recognition and the loss of watermarks. Watermarking algorithm based on natural language processing is to implement equivalent information replacement and word order modification on part of the text without changing the original meaning. This algorithm has two categories: one is based on syntax [Lei 2009]and the other is based on semantics [Atallah et al. 2001; Atallah et al. 2001].The former is to modify the Chinese characters themselves or to embed information after dividing the sentence into word senses [Atallah et al. 2002; Liu, Sun and Luo

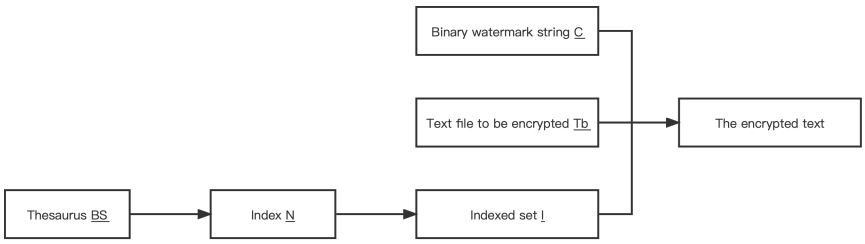

Figure 1: Schematic diagram.

2006]. The latter is to embed information after analyzing the words in the text [Bennett 2004].

Copying information is popular among users because science and technology have made it easier and more convenient. In recent years, optical character recognition (OCR) [Zhang et al. 2020] technology is widely used in extraction. Even when the security techniques-watermarking algorithm based on document structure and the watermarking algorithm based on traditional binary image are adopted, the watermark will disappear if the text content is extracted by OCR. Although natural language processing watermarking can resist the damage of watermark, it is deficient in the independence of water marks and fixity of elements to be embedded, which in turn leads to derivative problems such as low efficiency of watermarking calculation and low text reading speed.

This paper proposes a thesaurus-based method for generating and tracing unmarked text watermarking for the first time aiming at the deficiencies of the method. The main principle is to construct a fixed synonym thesaurus first. Next watermark based on the thesaurus will be generated and embedded. Then the method of tracing index based on the thesaurus will be worked out, including maintaining the independence between indexes to improve the coding accuracy, and controlling the number of indexes to improve the text reading speed. On this basis, this paper designs a robustness experiment and an invisibility experiment to verify the proposed methods.

2 GENERATION AND TRACEABILITY METHOD OF UNMARKED TEXT WATERMARKING BASED ON THESAURUS

2.1 General Idea

The main schematic diagram of this method is shown in Figure 1 below.

Firstly, a thesaurus is designed to generate a set of synonym pairs. On the basis of it, the elements are grouped based on specific requirements to generate an index set. Then the index set is used to cooperate with a given binary string to generate a text watermark and realize the watermark traceability.

2.2 Design and construction of thesaurus

2.2.1 Structure design of thesaurus. Compared with single characters, pairs can express more complete meanings in Chinese. The number of two-word or four-word pairs are greater. That's why the thesaurus designed in this research is composed of two-word synonyms and four-word synonyms and different pairs are separated by line breaks.

The embedded watermark is in binary form, which contains only 0 and 1, thus synonym pairs are endowed with numerical characteristics. In each group of synonym pair, the former represents logic 0 and the latter represents logic 1 so that the watermark binary digital information can match with text. The thesaurus data structure is shown in Table 1.

Thesaurus is the carrier library of embedded information and the core of watermark embedding. Once generated, it cannot be changed. When upgrading the thesaurus, its original version needs to be kept. Otherwise, the embedded information will be invalid.

2.2.2 Watermark embedding based on thesaurus. Based on this thesaurus, watermarks can be embedded. The practice is to divide the whole text into a set of phrases based on natural language processing. A repeatable array is formed according to the word order in the text and then the array of sequences is iterated over. Comparing the words obtained by each traversal in the thesaurus, if they match, synonym pairs will be replaced, then the corresponding digital information of the word in the thesaurus will the same as that to be encoded in binary code.

According to the embedding rules, the text results before and after embedding are compared as shown in Table 2 below.

2.3 Index generation and traceability method based on thesaurus

2.3.1 Index definition and generation method. Due to the large number of information carriers, numerous contents and unchangeable nature of thesaurus, frequent access to it will consume a lot of time and increase the risk of thesaurus changes. Therefore, when tracing watermarks, the thesaurus needs to be segmented into several independent sub-sets, i.e. indexes, which will be used to undertake the accessed work in a dispersed way. The process is shown in Figure 2 below.

There are many "specific rules" for index generation and different requirements for different index groups. But the basic requirement must be satisfied: words (in this case, not synonyms) between different indexes should not intersect in an index group. Otherwise, it will produce the ambiguity of logic and synonym matching.

2.3.2 Watermark embedding based on index. Similar to 2.2.2 Watermark embedding based on thesaurus, the watermarks will be

Table 1: Thesaurus structure and coding

logic 0	logic 1
Two - or four-word pairs A	A close synonym for the two-word or four-word pairs A
Two - or four-word pairs B	A close synonym for the two-word or four-word pairs B
Two - or four-word pairs C	A close synonym for the two-word or four-word pairs C
(Continue writing in the above format)	(Continue writing in the above format)
. . .. . .	. . .. . .

Table 2: Before and after embedding (Translated English Version)

Original document	The embedded document
Huazhong University of Science and Technology has elegant environment and beautiful scenery. Huazhong University of Science and Technology is known as "forest University". The school has a complete teaching and research support system and complete public service facilities .	Huazhong University of Science and Technology has elegant environment and great scenery. Huazhong University of Science and Technology is known as "forest University". The school has a thorough teaching and research support system and complete public service facilities .

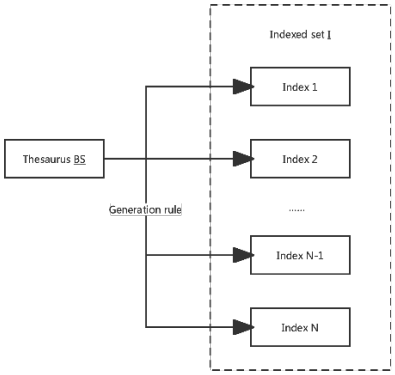

Figure 2: Index generate graph.

embedded based on the index. However, the index adopts specific rules for classification. When the word is located in the index, the search after locating a single index following the rules can save a lot of time.

2.3.3 Watermark tracing method. The watermark tracing method is the reverse process of 2.3.2 Watermark embedding based on thesaurus generation index. According to the text with embedded information, it is segmented into words, and a repeatable array is formed according to word order in the text. Then the array of sequences is iterated over. For the words matched in the index, the binary information referred to by them is recorded as an information bit of the binary coding. As the traversal of the information bit record gradually grows, the binary information is completed eventually.

2.3.4 Index efficiency improvement methods. In order to improve the efficiency of the index to the greatest extent, the index generation algorithm F is proposed.

Define the tuple(BS, N, I, F), where BS represents the thesaurus, N represents the number of target indexes, I represent the set of indexes, and F is the index generation algorithm.

The binary watermark string C to be embedded is introduced, and its generation principle will be described in detail in the process of 3.1. In order to ensure that the binary watermark string C is embedded efficiently, it should be written quickly and the number of characters occupied by the word encoding should be the shortest. Increasing the number of indexes N makes querying easier. Reducing the number of indexes N can reduce the loss caused by dictionary segmentation and improve the effective encoding of phrases. Therefore, the optimal number of indexes is the key to improve the efficiency of indexing algorithms. According to the use experience, five indexes of 1, 2, 4, 8 and 24 are selected under the premise of ensuring the discrimination. Suppose the text length is TN, the time required to embed a watermark is t, the length of the occupied text is n, and the total time is T. The relationship between the four is as follows $T = (TN \div n) \times t$. Change the number of indexes to make n and t as small as possible. Figure 4 below shows the values of the two dependent variables at different indexes.

Corollary 1: The relationship between indexes and write time (t) shows a negative correlation. The relationship between indexes and length of fields occupied by writing (n) is as follows: As the number of indexes increases, the length of fields occupied by writing (n) shows an upward trend, from which it can be concluded that the relationship between indexes and length of fields occupied by writing (n) shows a positive correlation.

Based on the conditions above, the optimal point of efficiency of binary watermark string C is figured out, and the scatter plot of length of fields occupied by writing (n) and the writing time (t) under different indexes are drawn. The relationship diagram is shown in Figure 5.

Figure 5 proves ***Corollary 1***. Since the length of fields occupied has no correlation with the writing time and the weight selection of the two variables after standardization shows strong subjectivity,

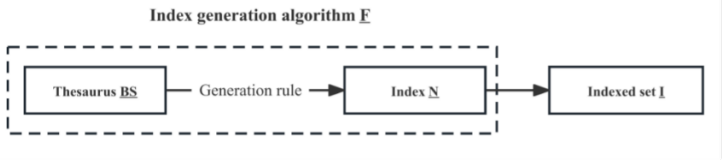

Figure 3: Index generation algorithm logic diagram.

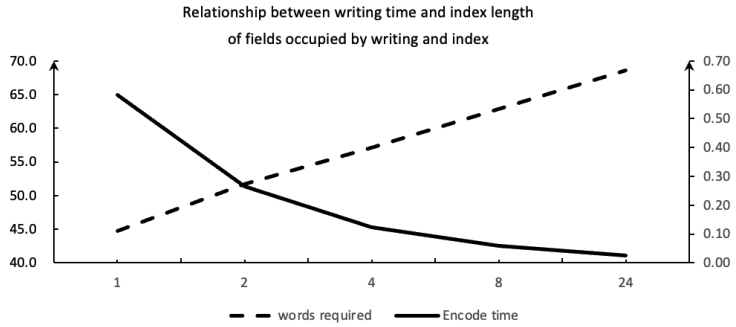

Figure 4: Relationship between writing time and index length of fields occupied by writing and index.

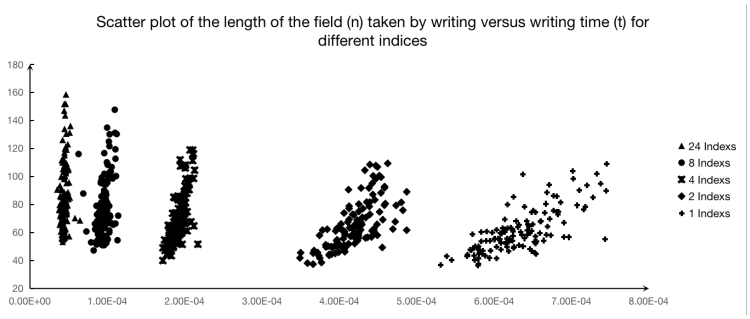

Figure 5: Scatter plot of the length of the field (n) taken by writing versus writing time (t) for different indices.

it is more reasonable to use qualitative judgment when selecting the number of indexes.

In the process of using indexes, time performance should be the primary consideration. In the mode of one and two indexes, the encoding time is long and fluctuated greatly, so it should be abandoned. Of the remaining indexes, four are more stable and should be selected.

The BS restriction given in the definition is as follows.

$$BS = \{(word11, word12) \ldots\ldots (wordn1, wordn2) \mid wordi1 \; stands \; for \; logical \; 0 \; \& \; wordi1 \; stands \; for \; logical \; 1 \& i \in n\} \tag{1}$$

The single index (I) property given the number of indexes N in the definition is as follows.

$$I = \{(word11, word12) \ldots\ldots (wordn1, wordn2) \mid (wordi1, wordi2) \in BS\& \; i \in n\& InitialLetter\,(wordi1) = InitialLetter\,(wordi2) \; Specific \; initial \; letter \; defined \; by \; I\} \tag{2}$$

From the definition of BS with I, the expression of F can be obtained.

$$F : Extract \; the \; initial \; letters \; from \; (wordi1, wordi2) \; and \; feed \; it \; into \; the \; qualified \; I \tag{3}$$

2.3.5 Approach to lexical adhesion problem. The automatic word segmentation method is used in text recognition. In the mode of natural language processing, words carrying information might disappear because of their adhesion to the context, and new words will be produced, which might bring about loss of information bits. We call this the “adhesion problem”, and it only occurs during embedding. When the problem arises, the missing information position should be repeatedly embedded to ensure the integrity of the coding.

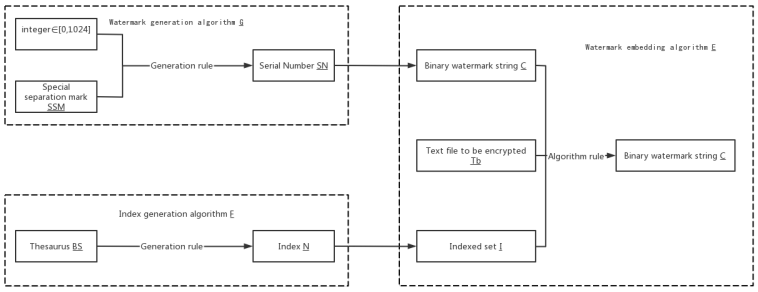

Figure 6: Watermark embedding process algorithm.

Table 3: Encoding attribute table

Coding type	Advantage	Disadvantage	Total num of codes (15)	Total num of codes (18)
CRC	Check for burst errors	The number of redundant bits is long	105	34
BCH	Strong error correction ability	Insufficient error detection capability, long redundancy	6	(No correspondence)
Parity check	Detecting continuous error	The error detection efficiency is lower than that of CRC	784	128

3 WATERMARK EMBEDDING AND TRACING ALGORITHM FLOW

3.1 Watermark embedding process and algorithm

According to tuple representation proposed in the 2.3.4 Index efficiency improvement methods, expand the tuples and define encryption tuple(SSM, SN, C, BS, N, C, Tb, I, C, G, F, E). BS represents thesaurus, N represents the number of target indexes, I represents the set of indexes, and F represents the generation algorithm of index. In this part of analysis, Tb represents the set of text to be encrypted, C represents the binary watermark string to be embedded, I represents the set of replacement reference indexes, G represents the watermark generation algorithm, F represents the dictionary generation algorithm. E represents the embedding algorithm of the watermark.

3.1.1 Watermark generation algorithm G module. When selecting the username number SN, the optimal code is selected among the parity check code, CRC cyclic redundancy check code, and BCH code.

The characteristics of each encoding are shown in Table 3. Because encoding requires more error detection than error correction (no one ever changes a synonym when modifying a text), parity checks that can detect continuous errors and encode a large number of codes are the best choice. Thus, the expression for SN is shown in the following expression (4).

$$SN = \left\{ w(B) \,|\, w(B) \in [0, 1023Byte] * 2^5 + check1 * 2^4 + check2 * 2^3 + B \right\} \quad (4)$$

The watermark generation algorithm G is used to subtract the special separated permutations from the total set SN.

$$C = \{ G(SN) \,|\, G(SN) \in (SN - redundant\ watermark) \} \quad (5)$$

$$redundant\ watermark \in \{ SN |\, (SN_i, SN_{i+1}, SN_{i+2}) = (B_0, B_1, B_2) \} \quad (6)$$

3.1.2 Watermark embedding algorithm E. Watermark embedding algorithm focuses on Chinese characters in text. English, numbers or symbols are not taken into consideration.

Based on the binary string C and index I, the watermark embedding algorithm E is used to embed the encrypted text Tb. The purpose is to iterate over the text TB and split it into a word set. In each traversal, if the obtained word (pending word) matches with index I, a bit Cj ($j \in length(C)$) of binary string C will be taken as a signal to decide what word can replace the matched one. The logic is expressed in the following expression (7).

$$E: \text{pending word} = \begin{cases} wordi1\ if\ Cj = 0 \\ wordi2\ if\ Cj = 1 \end{cases} when \text{ pending word} = \text{word}i1 \text{ or word}i2 \quad (7)$$

In the process of embedding binary string C, it is necessary to consider the problem of failure after replacement and define failure signals ($SubstantialResults$), that is, embedded parts will adhere to context and cannot be read by splitting. Bit needs to be re-embedded when the failure signal is 1.

$$SubstantialResults = \begin{cases} 1\ if\ pending\ word\ can't\ be\ split\ out\ after\ feed\ it\ into\ the\ article \\ 0\ if\ pending\ word\ can \quad be\ split\ out\ after\ feed\ it\ into\ the\ article \end{cases} \quad (8)$$

3.2 Watermark embedding process and algorithm

The watermark interpretation process is shown in the following figure, which is tuple (Tb, I, C, D).

3.2.1 Watermark extraction algorithm D module. Extracting module is the reverse process of embedded module logically. Based on

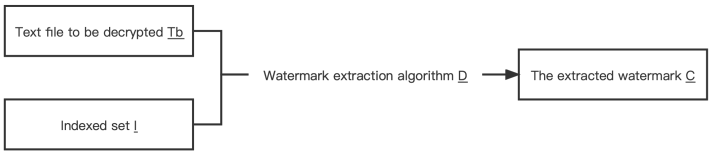

Figure 7: Watermark extraction process.

Table 4: Attack response effect table

Attack direction	watermarking survivability	correctly identifying ability	watermark crisis degree
Extractive attack	√	√	√
Multi-sentence added attack	√	√	√
Single-segment add attack	√	√	×
Periodic deletion attack	√	√	√
Characteristic delete attack	√	√	√
Substitution attack	√	√	√

index I, extracting binary watermark string C for decrypted text Tb is carried out. The decoded part has the same requirements for generating the index as the encoded part, so index I is required to be stable. The definition of decoding is as follows:

$$D : Cj = \begin{cases} 1 \; if \; processing \; word \; match \; a \; single \; index \; (I) \; \; and \; equal \; to \; \; wordi1, \; i \in n \\ 0 \; if \; processing \; word \; match \; a \; single \; index \; (I) \; \; and \; equal \; to \; \; wordi2, \; i \in n \\ else \; do \; nothing \end{cases} \tag{9}$$

When $C_i, C_{i+1}, C_{i+2} = (B_0, B_1, B_2)$, it means that a group of extraction has been completed, and groups of extraction run one loop after another until the end of the article.

4 EXPERIMENT AND ANALYSIS

Text watermarking should consider two characteristics: robustness and invisibility. Therefore, the experiment will be carried out in correspondence to the features mentioned above: (1) The robustness experiment of text watermarking (2) The invisibility experiment of text watermarking

4.1 The robustness experiment of text watermarking

Robustness testing is also called anti-attack testing. Since format attacks have no effect on text watermarking, the robustness test of text watermarking will focus on content attacks.

4.1.1 Experimental design. The following test method (a single experiment) is designed to verify the anti-aggression of traceable watermarking in unmarked text.

(1) Select a test file, use the traceable watermark in unmarked text, carry out watermark embedding of different binary string C on the article for many times, scan or identify the embedded article, and intercept the fragment.

(2) Sort and number the files embedded with watermark, and classify them according to the number. Attacks can be added, deleted, or changed aiming at the category.

(3) Watermark extraction is performed on the attacked files and then record the success rate of **watermark extraction**.

4.1.2 Experimental results and analysis. Based on the test method and attack type described in 4.1.1 Experimental design, the following results are obtained. The watermark survivability indicates whether the watermark still exists in the text. The correct recognition ability indicates whether the watermark can be correctly read. And the watermark crisis degree indicates that the watermark credibility ($Watermark\ reliability = Correct\ watermark \div All\ watermark$)is lower than 50% if the high-intensity attack is more than 50% on the article.

In Table 4, √ means that its contents are feasible, while × is just the opposite. The results show that the generative watermarking algorithm has strong resistance to text attacks. Under extreme attacks, it can still provide the normal recognition function and minimize the loss to ensure the reliability of the watermark, so it has strong robustness.

4.2 The invisibility experiment of text watermarking

Since good reading feeling is the ultimate goal of invisibility, the subjective feeling of subjects is used as the evaluation index in the invisibility experiment, and the subject's feedback on semantics is focused on.

4.2.1 Experimental design. The following test method is designed to verify the invisibility of tracing watermarking in generated unmarked text

(1) Select a test file, use tracing watermarking in generated unmarked text, carry out watermark embedding operation of different binary string C on the article for many times, scan or identify the embedded article, and intercept the fragment.

(2) Sort and number the files embedded with watermarks, record the actual number of watermarks under each number, and calculate

Table 5: Subject feedback

Subject class	0	1	2	3	4	5	6	7	8	9
Specific gravity qualification	√	√	√	√	√	√	×	×	√	√

the number of changes brought by the actual watermark embedding.

(3) Issue instructions to users and ask them to mark sentences or words that are not subjectively smooth.

(4) Count the actual number, calculate the overall proportion, and compare it with the expected proportion (no less than 85%).

4.2.2 Experimental results and analysis. For the feedback of ten different groups of subjects, the following results were obtained.

Table 5 shows that the hiding effect of this kind of watermark on users still needs to be improved, but the user's discovery of watermark has little influence on the watermark itself. Considering that the mainstream watermarking technology is still visible to users, the invisibility of watermarking is within the acceptable range.

5 CONCLUSION

This paper proposes a generative traceable watermarking technology in unmarked text. Articles can be cyclically embedded with watermarks and watermark tracing can be supported by means of natural language processing watermarking. In this paper, the generation of text watermark is completely set up, the generation and the decoding method of text watermark are also introduced in detail, and a complete idea and implementation path are provided as well. This paper focuses on the text watermarking cryptographic embedding, including embedding coding, the relationship between encoding and text replacement, coding segmentation and decoding anti-interference, which is significant to the generation of invisible watermarking. In particular, this paper makes deep research on the watermarking index and the value of index on improving the efficiency of watermarking. Experiments are carried out to prove the best index method and strong support for text watermarking is provided as well. In addition, this type of watermarking does not depend on the text format, so it will resist any type of format attacks effectively. As a result, it has theoretical and application value in synonym replacement, invisible text watermark embedding and watermark tracing.

ACKNOWLEDGMENTS

This research was completed under the guidance of Professor Ling Hefei, Director of Institute of Digital Media and Intelligent Technology, third-level professor, PhD supervisor, Huazhong Scholar, School of Computer Science and Technology, Huazhong University of Science and Technology. I would like to thank Professor Ling for his guiding opinions and help in this research. This research is also under the guidance of Professor Jiang Jie, Secretary General of International Society for Photogrammetry and Remote Sensing (ISPRS), Vice President of International Science Council of China Association for Science and Technology, Distinguished Dean of International Development Institute of Beijing University of Civil Engineering and Architecture (Institute of International Education), and second-level professor. I would like to thank Professor Jiang Jie for providing valuable guidance for the validation of this study and the interpretation of this paper. This research was guided by Lecturer Guo Xian of the School of Surveying and Mapping and Urban Spatial Information of Beijing University of Civil Engineering and Architecture. I would like to thank lecturer Guo Xian for his suggestions and plans for the modification of this paper. This research was done jointly with Yu Yifei, School of Computer Science and Technology, Huazhong University of Science and Technology, thanks for his contributions!

REFERENCES

[1] ATALLAH, M.J., MCDONOUGH, C.J., RASKIN, V. AND NIRENBURG, S. 2001. Natural language processing for information assurance and security: an overview and implementations. In *Proceedings of the 2000 workshop on New security paradigms* Association for Computing Machinery, New York, NY, United States, 51-65.

[2] ATALLAH, M.J., RASKIN, V., CROGAN, M., HEMPELMANN, C., KERSCHBAUM, F., MOHAMED, D. AND NAIK, S. 2001. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Information Hiding: 4th International Workshop* Springer, Pittsburgh, PA, USA, 185-200.

[3] ATALLAH, M.J., RASKIN, V., HEMPELMANN, C.F., KARAHAN, M. AND TRIEZENBERG, K.E. 2002. Natural Language Watermarking and Tamperproofing. *Springer Berlin Heidelberg*.

[4] BENNETT, K. 2004. Center for education and research in information assurance and security. *Linguistic Steganography 47906*.

[5] BRASSIL, J.T., LOW, S. AND MAXEMCHUK, N.F. 1999. Copyright protection for the electronic distribution of text documents. *Proceedings of the IEEE 87*.

[6] DING, H. AND HONG, Y. 2001. Interword distance changes represented by sine waves for watermarking text images. *IEEE Transactions on Circuits and Systems for Video Technology 11*, 1237-1245.

[7] GU, Y. AND FENG, J. 2015. A Digital Watermarking Algorithm for PDF Text Based on Space Encoding. *Journal of Foshan University of Science and Technology (Natural Science) 33*, 76-80+87.

[8] HUI, L. Research and implementation of digital watermarking algorithm based on WORD document. In *College of Computer Application Technology* Nanjing University of Science and Technology.

[9] KANG, S. 2008. Digital watermarking algorithm based on Chinese Word document. In *College of Computer Application Technology* Dalian University of Technology.

[10] LEI, L. 2009. A text watermarking algorithm based on natural language. *Journal of Guiyang University (Natural Science Edition) 4*, 39-43.

[11] LI, J. AND ZHU, X. 2008. Research on space coded digital watermarking algorithm. *Computer and Digital Engineering*, 104-106+113.

[12] LIU, Y., SUN, X. AND LUO, G. 2006. A new information hiding algorithm based on the structure of PDF document. *Computer engineering*, 230-232.

[13] LUO, Y., XU, X., ZHAO, Y., HE, L. AND FANG, D. 2008. Research on the Application of Text Watermarking in Digital Rights Management. *Microelectronics and computers*, 115-117.

[14] WU, M. AND LIU, B. 2004. Data hiding in binary image for authentication and annotation. *IEEE Trans. Multimedia 6*.

[15] XU, Z. 2013. Design and implementation of electronic document copyright protection system. In *School of Software Engineering* University of Electronic Science and Technology of China.

[16] ZHANG, C. 2007. Research on digital watermarking techniques for binary text images. In *School of Computer Systems Architecture* Chongqing University.

[17] ZHANG, T., MA, M. AND WANG, D. 2020. Research on OCR Character Recognition Technology. *Computer Technology and development 30*, 85-88.

[18] ZHAO, W., GUAN, H., HUANG, Y. AND ZHANG, S. 2020. Review of text watermarking technology. *Journal of Communication University of China (Natural Science Edition) 27*, 55-62.

[19] ZHAO, X., SUN, J. AND LI, L. 2008. Text watermarking algorithm based on character ladder edge adjustment. *Computer application 28*, 3175-3178+3182.