Research Article

A Novel Coverless Text Steganographic Algorithm Based on Polynomial Encryption

Bo Guan ¹, Lichun Gong,² and Yanzhao Shen ³

¹School of Electronics and Information Engineering, Ningbo University of Technology, Ningbo 315211, China

²Hangzhou Branch of China Telecom Co. LTD., Hangzhou 310000, China

³Shandong Institute of Blockchain, Jinan 250102, China

Correspondence should be addressed to Yanzhao Shen; shenyanzhao@sdibc.cn

Received 5 January 2022; Accepted 14 February 2022; Published 7 March 2022

Academic Editor: Zhen Wang

Copyright © 2022 Bo Guan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aiming at the problems of low text utilization rate and ambiguity of secret information extraction in the “tag + keyword” coverless information hiding methods, we propose a coverless information hiding method of Chinese text based on polynomial encryption in this paper. The secret data communication is realized by the analogously forged URLs transmission, which improves the security of mobile computing to avoid the risk that the carrier texts are attacked maliciously. The text utilization rate and retrieval success rate are improved by using tags to select multiple keywords to expand the number of keywords in the index table. The keyword's location in secret message is obtained through text vocabulary matching, and the tag index and location information are encrypted by polynomial for secret information transmission. The split keywords of secret message can be transmitted in disorder. The experimental result shows that the secret message can be successfully extracted by sorting the extracted keywords according to the frequency and relevance of their combined tags. In our experiments with different test databases, both the hiding success rate and the average hiding capacity are improved with higher test values than other existing methods.

1. Introduction

In recent years, smart terminal devices have been widely used. People can communicate with each other or send files in real-time by using some instant messaging apps, e.g., WhatsApp, Facebook Messenger, or WeChat. Most modern instant messaging applications have greatly facilitated human life communication. The researchers have paid more attentions on mobile computing, especially the security of mobile computing. The openness of Internet and the uncertainty of information dissemination have led to the frequent occurrence of citizen information leakage. According to the 46th China statistical report on Internet development issued by the China Internet Network Information Center (CNNIC)[1], the number of Internet users in China had reached 940 million by the end of June 2020. In these Internet users, there were 20.4% netizens who had suffered personal information leakage, 17% netizens who had suffered online fraudulent, and 9.9% netizens whose

user accounts with passwords had been stolen. In the report, the statistical number of the tampered websites in China was as high as 147,682 [1].

To ensure that information is not stolen or tampered with, the technologies of data encryption, intrusion detection, and information hiding are usually used to keep information secure and integrated [2]. For data encryption, there are various complex cryptography algorithms, which make the cipher text difficult to be cracked. However, the cipher text is always presented as some messy and unreadable codes, which is easy to attract the notice of attackers. Intrusion detection systems can prevent unauthorized users from accessing the system. However, the system information security of the legitimate users will be greatly harmed once the illegal attackers disguise themselves to invade the system and modify the data. Unlike data encryption and intrusion detection, the technology about information hiding or steganography is used for covert communication where the secret message is embedded into

the cover work available for mobile communication. Especially for steganography, the technology has high imperceptibility, which means that it is hard for attackers to detect or remove the secret message. Therefore, information hiding technology plays an important role in the research field of information security, involving secure data sharing and storage for mobile computing.

Conventional information hiding methods make use of the redundancy of digital carriers and the insensitivity of human vision to modify the carrier to achieve the purpose of embedding secret information. However, if the carrier is modified to carry some extra information, it is inevitable to leave traces, which means that the attackers can recognize or destroy the secret information by using some steganalysis tools. To solve the problem of low concealment and low security of the conventional information hiding methods, the coverless information hiding is proposed to share the secret information without modifying the carrier [3] between the sender and receiver. The secret message can be selected or generated from the shared carrier. Because of the characteristic of coverless, the carriers can resist most steganalysis methods and various malicious attacks.

According to the type of the digital cover, the steganography technology can be divided into four categories: image steganography, video steganography, audio steganography, and text steganography [4]. Among the four cover types, the text is frequently used in mobile communication with obvious advantages including small data amount, legibility, and fast transmission speed. Therefore, the research on text coverless steganographic algorithm has become a hot topic with high theoretical significance and practical value. According to whether the cover text is modified or not, the steganography algorithms can be divided into “traditional steganography algorithm” and “coverless steganography algorithm” [5]. The text coverless steganography can be further divided into two categories: the methods based on carrier generation [6, 7] and the methods based on carrier retrieval [8–10].

The steganography methods based on carrier generation need to establish a protocol about natural language text generation, which is shared by the two sides of communication. A steganographic text as the carrier is generated by the secret information control model. Fang et al. [6] used a new linguistic stegosystem based on a Long Short-Term Memory (LSTM) neural network to train language generation models, which achieved good concealment effects. Yang et al. [7, 11, 12] successively used Convolutional Neural Network (CNN), Markov Model (Markov Model), and Recurrent Neural Network (RNN) to extract text features and jointly trained the steganographic text generation model with Huffman coding.

The steganography methods based on carrier retrieval need to retrieve message from the cover text following some rules and communicate by transmitting the cover text. Chen et al. [8] proposed a conception of “tag + keyword” as the core idea for hiding information without carrier text modification, where the tag is calculated and related with Chinese mathematical expression. The method has high concealment, but low hiding capacity, where one character is

hidden in one text. Zhou et al. [13] improved the above method by combining multiple keywords hidden in one text to increase the hiding capacity. Liu et al. [14–16] used Chinese Pinyin and part of speech as tags and assigned keywords to each tag in a text. The proposed methods used language model of “Word2Vec” to expand the keyword set to increase the success probability of keyword retrieval. Zhang et al. [9] proposed the concept of “word rank map” to generate stegovectors for keyword retrieval. The method was improved to combine the frequent words hash against statistical analyzing attacks in [17] and combine the frequent words distance to increase utilization of retrieved text in [18]. However, the hiding capacity of the “word rank map” based methods in [9, 17, 18] is still low. In order to improve the coverless information hiding capacity, Ji et al. [10, 19] constructed a mapping relationship between binary strings and the number of keywords to realize multiple keywords hidden in one text.

The methods proposed above did not consider the distribution of tags in the text when selecting tags, which resulted in low word coverage under a single tag. Chen et al. [20, 21] considered the uniform distribution of the parity of Chinese characters’ Unicode encoding and designed a coverless information hiding method based on Unicode label. The method improved hiding success rate and hiding capacity, but it also needed a huge carrier text database to generate an index file, where the links for each tag are stored with only one selected keyword in a text. In addition, the transmitted secret message is directly hidden in the carrier text, which reduces the security of the method. Wang et al. [22] proposed a text coverless steganography method based on mathematical expression of Chinese character. Wang and Gao [23] further analyzed the parity of Chinese characters’ stroke number and its statistical characteristic and generated a communication codebook based on the exploited “space mapping” concept to further improve the embedding rate.

By analyzing the recent existing works, there are still some drawbacks in aspects of text utilization, information extraction accuracy, hiding capacity, and hiding success rate in the retrieval-based coverless steganography algorithms. In this paper, we propose a novel coverless steganography algorithm based on polynomial encryption for hiding secret message by Chinese text. It uses web page text as carrier and constructs an index file based on “tag + keyword” methods to collect the keywords location in carriers. The index file is shared with both sender and receiver for keywords retrieval. The novelty of the proposed algorithm includes two points: (1) corresponding one tag to multiple keywords and (2) generating retrieval side-information encrypted by polynomial for secret information transmission. The rest of this paper is organized as follows: in Section 2, the related works about encryption and decryption method, index file structure, and the framework of “tag + keyword” based steganographic scheme are reviewed. Section 3 presents our proposed coverless text steganographic scheme. Section 4 illustrates the feasibility of our proposed steganography by analyzing the experimental results with comparison of other “tag + keyword” methods. This is followed by conclusion and future work in Section 5.

2. Related Works

To present enough background knowledge for our proposed scheme, this section briefly reviews the general idea of secret sharing based on polynomial interpolation and the framework of previous steganography scheme based on “tag + keyword.”

2.1. Secret Sharing Scheme Based on Polynomial Interpolation.

A secret sharing scheme is to split a secret into n carry shares, which are used to reconstruct the original secret. Given a (k, n) threshold scheme and any shares less than k , nothing is learned about the secret. In our proposed steganography algorithm, Shamir’s secret sharing scheme with threshold $k = n$ is applied to encrypt our location side-information with various numbers of data. A dynamic key generation mechanism is established, where the number of keys is determined by the number of values in side-information.

2.1.1. Encryption Process Based on Polynomial. The n -degree polynomial used for encryption is defined in

$$f(x) = (a_0 + a_1x^1 + a_2x^2 + a_3x^3 + \dots + a_{n-1}x^{n-1} + a_nx^n) \bmod q. \quad (1)$$

In the encryption process, the plain text is the side-information with n integer values $P = [p_1, p_2, \dots, p_n]$, which are used to replace the coefficients of the above polynomial. A set of n keys $X = [x_1, x_2, \dots, x_n]$ is generated by a secret value $s \in [0, q)$ under prime power q . And the first coefficient of polynomial is set to be $a_0 = s$. Then, the polynomial is constructed as follows:

$$f(x) = (s + p_1x^1 + p_2x^2 + p_3x^3 + \dots + p_{n-1}x^{n-1} + p_nx^n) \bmod q. \quad (2)$$

Finally, the secret shares as output $\tilde{P} = [\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_n]$ are calculated as the corresponding value of $f(x_i)$ calculated by Equation (2).

2.1.2. Decryption Process Using Lagrange Polynomials for Polynomial Interpolation. When the receiver gets the secret shares as cipher text, n equations are formed as follows:

$$\tilde{p}_i = (s + p_1x_i^1 + p_2x_i^2 + p_3x_i^3 + \dots + p_{n-1}x_i^{n-1} + p_nx_i^n) \bmod q. \quad (3)$$

The plain text, n unknowns coefficients $p_1, p_2, \dots, p_n$, can be decrypted using Lagrange polynomial interpolation method. The Lagrange polynomial is defined in (4), where the corresponding coefficients of calculated $L(x)$ are the output plain text $P = [p_1, p_2, \dots, p_n]$.

$$L(x) = \sum_{j=0}^n \tilde{p}_j \prod_{i=0, i \neq j}^n \frac{x - x_i}{x_j - x_i}. \quad (4)$$

2.2. The Framework of General Coverless Steganographic Schemes Based on “tag + keyword”. The text coverless Steganographic schemes based on “tag + keyword” utilize the

natural language processing technology to cut the secret message into different keywords and use big data retrieval technology to retrieve the stegotext. The general framework of “tag + keyword” based method is shown in Figure 1. The implementation process includes four main steps: building a large text database, generating index files, collecting the stegotexts, and extracting hidden information.

Step 1. Building a large text database

Using web crawler technology, the text database can be quickly and accurately built by collecting different fields texts from different websites. The steganography scheme based on “tags + keywords” has to build a huge database with size of more than 10 GB in order to keep a higher hiding efficiency. They also analyze the frequency of keywords and their combinations of each text, which can be used to select carrier texts more efficiently and scientifically.

Step 2. Generating index files

The index file plays a crucial role in keeping the efficient retrieval. The index structure of the index file is inverted in the process of its creation process. There are three levels in index structure. The first level is *tag*, which is used to locate the secret message word. All the texts in database are pre-processed, and the location tag is selected and retrieved according to the binary strings corresponding to the parity of character’s Unicode. The first time for each location tag appearing in one text will be recorded with its keywords in the second level of index file and its “text path + file name” in the third level of index file. After all the texts are processed, the index file is finally created.

Step 3. Collecting the stegotexts

In this process, the secret message is firstly split into keywords. For each keyword, a corresponding tag is combined with it, and the “tags + keywords” are retrieved in index file to obtain the available texts’ paths. Select one text randomly and add it into the set of stegotexts, which will be sent to the receiver in sequence.

Step 4. Extracting hidden information

The process of secret information extraction is opposite Step 3. When the receiver gets all the stegotexts, the corresponding tags are combined with the texts in sequence. For each text, the keyword can be extracted according to the corresponding location tag. After all the received texts have been processed, the secret message can be constructed by combining all extracted keywords in the correct sequence.

3. The Proposed Coverless Text Steganographic Scheme

3.1. Research Motivation. Although the existing methods based on “tag + keyword” have fulfilled the basic requirements for coverless text information hiding, there are still two main issues about their disadvantages worthy of attention.

FIGURE 1: The framework of general coverless steganographic schemes based on "tags + keywords."

3.1.1. Lower Text Utilization. In order to ensure that the recipient can correctly extract secret keywords under the guidance of tags from the texts, one tag is only allowed to combine with one keyword in a fixed text during the stage of index file generation. It is possible that each tag appears multiple times in the following content of the fixed text and corresponds to multiple different keywords. In consequence, if there is more than one time for a tag appearance in the text, only the first appearance with its corresponding keyword will be kept and stored in the index file. In order to achieve a higher retrieval success rate, the capacity of text database needs to reach tens of GB or even hundreds of GB if the method based on "tag + keyword" is applied by using one-to-one correspondence between tag and keyword. Furthermore, the experimental studies show that the occurrence possibility of one tag appearing multiple times in a text is extremely high. That is bound to ignore much more utilizable keywords in one text. And the utilization rate for one text is not high by using one-to-one "tag + keyword" based method.

3.1.2. Ambiguity in Secret Information Extraction Caused by Sending and Receiving Texts Nonsequentially. The protocol of location-tags generation is agreed upon in advance by both communicating parties, which means that the order of the tags and their corresponding cover texts are fixed. If the sender does not send the texts in the order of secret information retrieval, or the receiver does not extract the secret information in correct order, or the texts transformation is affected by the channel delay, which results in receiving texts in disorder, all above cases can make the texts disordered, and the secret message extracted corresponding to ordered tags can make the information ambiguous.

To solve the problems mentioned above, we propose a novel coverless steganographic method based on polynomial encryption for Chinese characters. The testing text database was constructed by collecting texts on different web-pages. The uniform resource locator (URL) of the web page instead of web page text was stored in an index file. The polynomial coefficients are replaced by values calculated from all

keywords with their corresponding tags and location index in one text. The polynomial encryption algorithm is applied to generate secret information as auxiliary information, which is attached to the end of the text URL. Then, the new counterfeit URL will be sent to the receiver to realize the secret communication between two parties. In this method, one text can be utilized for hiding one or more than one secret keywords. Tags are not only used to localize the hidden keywords, but also used to improve the randomness of the algorithm, thereby improving the security of the algorithm. In addition, the proposed method uses Unicode Standard for character encoding, and the parity of encoded Chinese characters is utilized for tag retrieval. It has been proved in the literature [20, 21] that the distribution of parity for the encoded web page texts is uniform, which provides a guarantee for uniformity of the generated index file and a high word coverage ability.

3.2. The Framework of Proposed Method. The overall framework of our proposed coverless steganographic method based on polynomial encryption is shown in Figure 2. As shown in Figure 2, the proposed the method is mainly composed of three stages: *index file generation*, *secret information hiding*, and *secret information extraction*, which will be introduced in detail in Section 3.2.1, 3.2.2, and 3.2.3, respectively.

3.2.1. Index File Generation. The index file generation is primarily prepared for the subsequent quick retrieval of secret information. The main steps in the process of index file generation include generating a tag-set, initializing an index structure, preprocessing the web page texts, selecting the keywords according to each tag, and adding the keywords and their corresponding web page links in the index table. Algorithm 1 describes the procedure of the index file generation in details, and Figure 3 shows a text processing example for keywords selection and index table expanding with the tag-length $m = 4$. The structure of generated index file is illustrated in Figure 4.

FIGURE 2: The framework of proposed coverless steganographic schemes.

FIGURE 3: Conversion mode for index file generation.

FIGURE 4: The internal structure of the index file.

In the process of lines 6–14 in Algorithm 1, one tag with a specific value may appear more than one time for each one preprocessed text t_i . As long as the keywords following the

specific tag are not repeated, which means that the “tag + keyword” is unique, the searched keywords and their web page link will also be stored in the index structure under

Input: web page text database $T = [t_0, t_1, t_2, \dots, t_{n-1}]$ and length of tags m

Output: index file

- 1: Generate the tag-set according to the length of tags m . There are totally 2^m tags in a tag-set, each of which is a unique binary string with m -bit length.
- 2: Initialize an inverted index structure for each tag. The index table is a three-level tree structure, where the first-level index is for tags, the second-level index is for keywords, and the third-level index is for web page links.
- 3: **for** each text $t_i, i = 0, 1, 2, \dots, n-1$ **do**
- 4: Apply the preprocessing operations on each text t_i in web page text database to remove all non-Chinese characters (e.g. numbers, mathematical letters, punctuation marks), and finally the preprocessed text $\tilde{t}_i$ with only Chinese characters is generated.
- 5: Each Chinese character is converted into 0 or 1 according to the parity of its Unicode encoding value. For each preprocessed text $\tilde{t}_i$, a binary string S is generated according to the parity of character encoding.
- 6: **for** $j = m, m+1, m+2, \dots, \text{length}(S)$ **do**
- 7: Select the m -bit length binary string as the tag, $\text{tag} = S(j-m : j-1)$
- 8: Select four characters in $\tilde{t}_i$ following the location index j .
- 9: Split the 4 characters into different keywords.
- 10: The first keyword is chosen to be combined with the tag.
- 11: **if** “tag + keyword” is unique **then**
- 12: The “tag + keyword” information details (including the web page link of t_i) are added into the inverted index structure to expand the index file.
- 13: **end if**
- 14: **end for**
- 15: **end for**

ALGORITHM 1: Index file generation.

their corresponding tag branch. After all texts in the web page text database have been processed, the index file will be generated by merging 2^m tag-index branches.

3.2.2. Secret Information Concealing. In the process of secret information concealing, the secret message is firstly split into multiple keywords. All the keywords make up a Message-Keyword Set (S_{MK}). In the loop of processing “tag + keyword” retrieval, it is possible that some specific keyword, which is failed for once retrieval, is further split into smaller keywords. Meanwhile, the set S_{MK} is updated, where the failed retrieved keyword is replaced by more smaller keywords with less Chinese characters. Then, the split keywords and the remaining keywords in S_{MK} continue to do the retrieval loop one by one, and S_{MK} is continually updated. The retrieval loop continues until the keyword cannot be split further and there is no remaining keywords in S_{MK} .

Index File is used for each time of retrieval. For each successful retrieval time in the loop, the retrieved “tag + keyword” and its corresponding web page links are recorded. When the loop is finished, all the recorded data sets are used to count up how many keywords in S_{MK} appeared in every carrier text. The text, which has the highest number of keywords, is firstly chosen to extract location information with two location marks for each keyword. All the location marks are selected to make up a Side-Information set (S_{SI}) and to replace the polynomial coefficients for applying polynomial encryption once. Then, the set with encrypted marks S_{EM} is generated, and the encrypted marks as the confidential information are attached to the end of web page URL by some particular rules. After removing all the processed keywords in S_{MK} , the text which has the next highest number of remaining keywords is chosen to repeat

the same processing steps mentioned above. Finally, all the analogously forged URLs will be sent to the receiver for secret information transformation. The detailed procedure of secret information concealing is described in Algorithm 2.

3.2.3. Secret Information Extraction. The process of secret information extraction is opposite to the process of secret information concealing. For each received URL, the recipient firstly split it into parts: the web page URL and the confidential information part. The carrier text is downloaded from the web page URL and preprocessed before keywords extraction. The confidential information can be decrypted to obtain the side-information by applying Lagrangian interpolation. Then, we can extract the keyword(s) from the carrier text with side-information. It is possible that the secret message can be restored accurately, even if the sender did not send web page links in order or the receiver received the web page links in disorder. The whole extraction process works as follows:

Step 1: according to the agreed protocol, a disordered tag-set $L = [l_0, l_1, l_2, \dots, l_{2^m-1}]$ is generated firstly.

Step 2: when the receiver receives one or more URL messages, each URL is split into two parts: the web-page link and the confidential information part. From the web-page link, the corresponding carrier text is downloaded. From the confidential information part, the number of encrypted marks and the specific data for each mark can be analyzed according to the agreed rules. All the encrypted marks, which are calculated from one URL, constitute a set S_{EM} . And we can judge the number of keywords hidden in the text of a web-page based on the number of elements in S_{EM} .

Input: secret message and index file

Output: analogously forged URLs

- 1: Generate the disordered tag-set $L = [l_0, l_1, l_2, \dots, l_{2^m-1}]$ according to the protocol negotiated by the communication parties. m is the length of tag.
- 2: Apply preprocessing to the secret message to remove any non-Chinese characters, and then use the word segmentation tool to split the secret message into multiple keywords $S_{MK} = [k_0, k_1, k_2, \dots, k_{num-1}]$, where num is the number of keywords in S_{MK} .
- 3: $i = 0$ and $j = 0$
- 4: **while** $j < num$ **do** $\setminus(\triangleright \setminus)$ This is the process of retrieval for each split keyword in S_{MK} .
- 5: Select $l_i + k_j$ as the “tag + keyword” to apply retrieval in the index file.
- 6: **if** successful retrieval **then**
- 7: Return the set of carrier text URLs $U_i = [u_0, u_1, u_2, \dots]$ corresponding to $l_i + k_j$
- 8: $j++$
- 9: $i = \text{mod}(j, 2^m)$
- 10: **else if** The keyword k_j 's retrieval is failed in tag l_i branch of index file, and the number of characters in keyword k_j is more than 1. **then**
- 11: The keyword k_j is further split into smaller keywords $k_{j1}, k_{j2}, \dots$
- 12: The set S_{MK} is updated, where k_j is replaced by $k_{j1}, k_{j2}, \dots$
- 13: num is recalculated. and Let $k_j = k_{j1}$.
- 14: **else**
- 15: The retrieval for k_j is failed.
- 16: **end if**
- 17: **end while**
- 18: The returned U_j sets corresponding to all successfully retrieved keywords are used to count up the occurrence frequency frq of keywords in every carrier text.
- 19: The text $\tilde{t}_x$ which has the highest value of frq is selected.
- 20: **while** $frq \geq 2$ **do**
- 21: **for** each k_j in $\tilde{t}_x$ with its corresponding tag l_i **do**
- 22: Record two location marks in the side-information set S_{SI} . One is the location of k_j appeared in the word-list which contains all the different words following tag l_i in $\tilde{t}_x$. The other is the location of l_i in L .
- 23: **end for**
- 24: Apply polynomial encryption for S_{SI} to generate the encrypted marks S_{EM} .
- 25: Forge an analogous URL with confidential information in S_{EM} .
- 26: After removing all the processed keywords in S_{MK} , $\tilde{t}_x$ with the highest value of renewed frq is selected again.
- 27: **end while**
- 28: **for** each remaining keyword k_j appeared in one or more texts with $frq = 1$ **do**
- 29: Randomly select a URL in U_j to obtain $\tilde{t}_x$.
- 30: Record two location marks in S_{SI} .
- 31: Apply polynomial encryption with S_{SI} as input to generate S_{EM} .
- 32: Forge an analogous URL with confidential information.
- 33: **end for**
- 34: All the analogously forged URLs will be sent to the receiver for secret information transformation.

ALGORITHM 2: Secret information hiding.

Step 3: according to the initial value, step size, known prime number, and S_{EM} , the polynomial is reconstructed by applying the Lagrangian interpolation algorithm. The polynomial coefficients can be calculated and constitute the side-information set S_{SI} as output of decryption algorithm.

Step 4: obtain the preprocessed text by deleting all non-Chinese characters, and convert all Chinese characters into a binary string according to the parity of Unicode encoding. The number of secret message keywords hidden in the text may be one or more. For each keyword retrieval, the corresponding tag value is selected according to the tag position index storing in S_{SI} . The binary string is traversed to search the tag and generate a word-list that contains all the text words following the tag. Suppose that the tag is l_i and its

corresponding keyword is k_j . From the side information, the location of k_j appeared in the word-list is known and used to extract the keyword.

Step 5: after all web-page links are processed according to Steps 1–4, the extracted keywords are filled in the corresponding positions of secret message according to the tag position index. And finally, the secret message is obtained by correct permutation and combination of all keywords.

4. Experimental Results and Analysis

The testing database was constructed by 10.28 GB data of web page texts, which were collected from the Internet by crawler tool written in Python. The content of web page texts covers various aspects, including ancient Chinese poems,

novels, and news and knowledge introduction. A number of experiments in this paper are implemented to evaluate the performance of the proposed steganographic algorithm and compared with different existing algorithms in aspects of hiding success rate and average hiding capacity, respectively.

4.1. Selections of Parameters. Taking the size of 1 GB as the gradient, we totally construct five testing databases with 1 GB, 2 GB, 3 GB, 4 GB, and 5 GB of web page data with texts randomly selected from different topical contents. For each web page text database, the index file is generated by setting the tag length $m = 4, 5, 6$, respectively. The size values of index files, which are generated by different databases with different tag length settings, are shown in Table 1. As can be seen from the table, the index file size increases with both the tag length and the database size. Since the index file stores keywords and web page links, the size of the index file indirectly reflects the success rate of hiding. A larger index file contains more keywords and links inside, which makes the possibility of successful retrieval much higher for hiding information.

The secret message for information hiding is collected and stored in four sets. The experiments for performance comparison are implemented on the four message sets (MS) named as “MS-50-1,” “MS-50-2,” “MS-100-1,” and “MS-100-2,” respectively. The first number (50 or 100) in the name means how many secret messages are contained in the set. The second number (1 or 2) in the name means that the average size of the messages is 1 KB or 2 KB.

4.2. Success Rate of Hiding. The success rate (SR) of hiding, also known as retrieval success rate, is used to evaluate the quality of the extracted secret message. It is defined by (5) mathematically.

$$SR = \frac{\text{the number of successfully hidden messages}}{\text{the total number of secret messages}}. \quad (5)$$

In this paper, the previous “tag + keyword” method proposed by Chen et al. [20] is simulated and compared with our proposed method in different experimental settings. In terms of the effectiveness of information hiding, our proposed steganography algorithm has two differences in process design compared with the previous “tag + keyword” method. Firstly, in the index file generation process, multiple keywords can be selected for one tag in a text. Even if a relatively smaller web page text database is used, each tag in the index file can correlate with much more keywords or characters, which are enough for keywords retrieval in one secret message. Therefore, each word in the text can be fully utilized, and the probability of the secret keyword being retrieved can be increased. Secondly, a polynomial encryption algorithm is used to encrypt the location information and tag index, which is appended to the tail of URLs for transmission. The design of location information and tag index as side-information solves the problem that the

disordered data sending and receiving may easily cause the ambiguity of secret message extraction.

Figure 5 shows the hiding SR by applying the two simulated methods with different cover text databases and secret message sets. Here, the length of location tag m is set to be 6. It can be seen from Figure 5 that the hiding success rate of our proposed method can nearly approach 100% for any size of web page text database. For both methods, the hiding success rate goes up with the increased size of web page text database. It is obvious that even if the smallest 1 GB database is selected for testing, our proposed method can approach above 95% success rate, which is much higher than the method of Chen et al. [20]. In addition, the hiding success rate hardly changes when the size of the secret message increases from 1 KB to 2 KB and the number of messages increases from 50 to 100, which means that the stability of our proposed method performs good.

4.3. Average Hiding Capacity. Hiding capacity (HC) refers to the length of the secret message (i.e., the total number of Chinese characters in the secret message) hidden in each one carrier text. It is defined by Eq. (6) mathematically.

$$HC = \frac{\text{the length of the secret message}}{\text{the total number of the carrier texts}}. \quad (6)$$

The average hiding capacity (AHC) is calculated by

$$AHC = \frac{\sum_{i=1}^N HC}{N}, \quad N = \text{the total number of secret messages}. \quad (7)$$

The performance of AHC for Chen et al.’s method [20] and our proposed method is illustrated in Figure 6. There are four groups of test values, which show the comparison AHC values with experiments implemented on the four message sets: “MS-50-1,” “MS-50-2,” “MS-100-1,” and “MS-100-2.” Here, the length of location tag m is set to be 6.

It can be seen from Figure 6 that our proposed method has a higher hiding capacity under the same size of web page text database. This is due to the index file generated by our proposed method, which contains a large number of keywords with longer word-length (i.e., more characters). Therefore, the possibility of the keyword, which will be split into smaller segments, is reduced. And fewer carrier texts can be used for information hiding of a whole secret message, which leads to an increase in the average hiding capacity.

In addition, in order to demonstrate the improvement of our proposed method, there are more “tag + keyword” methods [8, 19–21], which are simulated and implemented to evaluate the performance of AHC. Table 2 lists the comparison values of AHC with different “tag + keyword” methods applied to four sets of secret messages. Here, we use the 1 GB web page text database, and the length of location tag m is set to be 6. It can be seen from Table 2 that the average hidden capacity of our proposed method is significantly higher than the existing similar methods proposed in [8, 19–21].

TABLE 1: The size of index files.

Tag length	1 GB database	2 GB database	3 GB database	4 GB database	5 GB database
$m = 4$	1.93 GB	3.75 GB	5.46 GB	7.12 GB	8.90 GB
$m = 5$	2.16 GB	4.18 GB	6.07 GB	7.90 GB	9.87 GB
$m = 6$	2.42 GB	4.66 GB	6.74 GB	8.15 GB	10.9 GB

FIGURE 5: Comparison of hiding success rate.

4.4. Security Analysis. In terms of the security performance, the improvement of our proposed method is demonstrated in four aspects. Firstly, the transmission data is the URLs, which contain the links for cover texts downloading and the

side information for message keywords extraction. In the process of data transmission, the carrier texts are neither modified nor directly transferred to the receiver, which can reduce the possibility of the risk that the carrier text is

FIGURE 6: Comparison of average hiding capacity.

TABLE 2: Comparison of average hiding capacity by using different steganography algorithms.

Methods	MS-50-1	MS-50-2	MS-100-1	MS-100-2
Method [8]	1.00	1.00	1.00	1.00
Method [19]	2.32	2.27	2.21	2.23
Method [20]	1.73	1.60	1.60	1.57
Method [21]	2.13	2.00	2.10	2.01
Our proposed method	2.79	3.82	3.04	3.82

attacked by malicious attackers. Secondly, the randomness of tag selection combining with the message keywords improves the security of our method. For example, the retrieved sets of web page links may be different by using

different tags to retrieve the same keywords, and the keywords selected in the same text by different tags are also different. Thirdly, the number of location information and the tag index for once polynomial encryption with a specific

key is dynamical. It is hard for attackers to guess the encryption process with different keywords splits and combinations. Finally, it is hard for the attacker to reconstruct the corresponding polynomial when the initial value, step size, and prime number are unknown, which means that the side-information decryption is rarely possible.

5. Conclusion and Future Work

For the existing “tag + keyword” coverless steganography methods, there are some obvious drawbacks in text utilization rate, ambiguity of secret information extraction, and hiding capacity. In this paper, we propose a coverless steganography algorithm of Chinese text based on polynomial encryption. The novelty of our proposed method contains two main points. Firstly, in the index file generation process, one tag corresponded to multiple keywords in a carrier text. Therefore, the effectiveness of information hiding in aspects of fully word utilization, and the probability of the secret keyword retrieval can be increased. Secondly, a polynomial encryption algorithm is used to encrypt the location information and tag index, which is appended to the tail of URLs for message transmission. The design of location information and tag index as side-information solves the problem that the disordered data sending and receiving may easily cause the ambiguity of secret message extraction. The experimental result shows that the secret message can be successfully extracted by sorting the extracted keywords according to the frequency and relevance of their combined tags. Compared with the previous “tag + keyword” based methods, our proposed method has a higher hiding success rate and the average hiding capacity in our experimental results.

In future research, there are two main issues to explore. One is to analyze the stability of the proposed data hiding scheme with more realistic settings, such as upgrading the website cover text. The other is to extend our proposed data hiding scheme to not only Chinese language, but also other languages with similar writing system, e.g., Japanese, Mongolic, and Korean languages. The basic principle of our proposed method can be applied to many other languages with similar keyword character-feature that all sentences can be divided into different words and all the words consisted with fewer characters (normally less than four). Future studies could investigate the feasibility of the data hiding scheme for different languages.

Data Availability

The text data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Zhejiang Provincial Natural Science Foundation of China (Grant Nos. LQ20F020019 and LY21F020007).

References

- [1] China Internet Network Information Center, “The 46th China statistical report on internet development,” *Journal of the National Library of China*, vol. 1, pp. 1–89, 2021.
- [2] E. E. Schultz, R. W. Proctor, M.-C. Lien, and G. Salvendy, “Usability and security an appraisal of usability issues in information security methods,” *Computers & Security*, vol. 20, no. 7, pp. 620–634, 2001.
- [3] Z. Zhou, Yi Cao, and X. Sun, “Bag-of-words model-based image-free information hiding,” *Journal of Applied Sciences*, vol. 34, no. 5, pp. 527–536, 2016.
- [4] N. S. Kamaruddin, A. Kamsin, L. Y. Por, and H. Rahman, “A review of text watermarking: theory, methods, and applications,” *IEEE Access*, vol. 6, pp. 8011–8028, 2018.
- [5] Z. Zhou, H. Sun, R. Harit, X. Chen, and X. Sun, “Coverless image steganography without embedding, Cloud Computing and Security,” in *Proceedings of the International Conference on Cloud Computing and Security*, pp. 123–132, Springer, 2015.
- [6] T. Fang, J. Martin, and K. Argyraki, “Generating steganographic text with LSTMs,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-Student Research Workshop*, pp. 100–106, Vancouver, Canada, July 2017.
- [7] Z. Yang, W. Nan, J. Sheng, Y. Huang, and Yu-J. Zhang, “TS-CNN: Text Steganalysis from Semantic Space Based on Convolutional Neural Network,” *Application Of Electronic Technique*, <https://arxiv.org/abs/1810.08136>, 2018.
- [8] X. Chen, H. Sun, Y. Tobe, Z. Zhou, and X. Sun, “Coverless information hiding method based on the Chinese mathematical expression,” in *Proceedings of the International Conference on Cloud Computing and Security, ICCCS 2015*, pp. 133–143, Springer, Nanjing, China, August 2015.
- [9] J. Zhang, J. Shen, L. Wang, and H. Lin, “Coverless text information hiding method based on the word rank map,” *Cloud Computing and Security*, Springer, in *Proceedings of the International Conference on Cloud Computing and Security*, August 2015.
- [10] H. Ji and Z. Fu, “Coverless information hiding method based on the keyword,” *International Journal of High Performance Computing and Networking*, vol. 14, no. 1, pp. 1–7, 2019.
- [11] Z. Yang, S. Jin, Y. Huang, Y. Zhang, and H. Li, *Automatically generate steganographic text based on Markov model and Huffman Coding*, <https://arxiv.org/abs/1811.04720>, 2018.
- [12] Z. Yang, X. Guo, Z. Chen, Y. Huang, and Y. Zhang, “RNN-Stega: linguistic steganography based on recurrent neural networks,” *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1280–1295, 2018.
- [13] Z. Zhou, Y. Mu, C.-N. Yang, and N. Zhao, “Coverless multi-keywords information hiding method based on text,” *International Journal of Security and its Applications*, vol. 10, no. 9, pp. 309–320, 2016.
- [14] Y. Liu, J. Wu, and G. Xin, “Multi-keywords carrier-free text steganography based on part of speech tagging,” in *Proceedings of the 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 2102–2107, IEEE, Guilin, China, July 2017.
- [15] Y. Liu, J. Wu, and G. Xin, “Multi-keywords carrier-free text steganography method based on Chinese pinyin,” *International Journal of Computational Science and Engineering*, vol. 21, no. 2, pp. 202–209, 2020.

- [16] Yi Long, Y. Liu, Y. Zhang, X. Ba, and J. Qin, "Coverless information hiding method based on web text," *IEEE Access*, vol. 7, Article ID 31926, 2019.
- [17] J. Zhang, H. Huang, L. Wang, H. Lin, and D. Gao, "Coverless text information hiding method using the frequent words hash," *International Journal on Network Security*, vol. 19, no. 6, pp. 1016–1023, 2017.
- [18] J. Zhang, Y. Xie, L. Wang, and H. Lin, "Coverless text information hiding method using the frequent words distance," *Cloud Computing and Security*, in *Proceedings of the International Conference on Cloud Computing and Security ICCCS 2017*, pp. 121–132, Springer, Nanjing, China, June 2017.
- [19] H. Ji, Z. Fu, and X. Sun, "Coverless information hiding method based on the label model," in *Proceedings of the 3rd International Conference on Cloud Computing and Security*, vol. 19, pp. 1509–1514, Nanjing, China, January 2018.
- [20] X. Chen, H. Sun, Y. Tobe, Z. Zhou, and X. Sun, "Coverless information hiding method based on the Chinese character encoding," *Journal of Internet Technology*, vol. 18, no. 2, pp. 91–98, 2017.
- [21] X. Chen and S. Chen, "Text coverless information hiding based on compound and selection of words," *Soft Computing*, vol. 23, no. 15, pp. 6323–6330, 2019.
- [22] J. Wang, Z. Guo, and K. Wang, "A method of text steganography without carrier based on mathematical expressions of Chinese characters," *Journal of Qingdao University (Natural Science Edition)*, vol. 32, no. 1, pp. 81–86, 2019.
- [23] K. Wang and Q. Gao, "A coverless plain text steganography based on character features," *IEEE Access*, vol. 7, pp. 95665–95676, 2019.