Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.017720
Article

Tech Science Press

A Robust Text Coverless Information Hiding Based on Multi-Index Method

Lin Xiang[1], Jiaohua Qin[1,*], Xuyu Xiang[1], Yun Tan[1] and Neal N. Xiong[2]

[1]College of Computer Science and Information Technology, Central South University of Forestry and Technology, Changsha, 410004, China
[2]Department of mathematics and computer science, Northeastern State University, OK, 74464, USA
*Corresponding Author: Jiaohua Qin. Email: qinjiaohua@163.com
Received: 08 February 2021; Accepted: 16 April 2021

Abstract: Recently, researchers have shown that coverless information hiding technology can effectively resist the existing steganalysis tools. However, the robustness of existing coverless text information hiding methods is generally poor. To solve this problem, we propose a robust text coverless information hiding method based on multi-index. Firstly, the sender segment the secret information into several keywords. Secondly, we transform keywords into keyword IDs by the word index table and introduce a random increment factor to control. Then, search all texts containing the keyword ID in the big data text, and use the robust text search algorithm to find multiple texts. Finally, these texts are converted into mixed indexes sent to the receiver. The receiver disassembles received indexes through the index construction protocol and uses the random increment factor to extract the secret information. Experimental results show that this method improves the concealment and security of secret information and has strong robustness compared with the state-of-the-art methods.

Keywords: Coverless information hiding; random increment factor; multi- index; text coverless

1 Introduction

Coverless information hiding has a long history, while Chinese ancient acrostic poetry is one of the most direct application cases. In recent years, with the rapid development of deep learning, natural language processing, and other technologies, coverless information hiding has rapidly developed, widely used in images, videos, etc. Luo et al. [1] proposed coverless image steganography based on Multi-object recognition, which has considerable hiding capacity and hiding rate. Qin et al. [2] proposed a coverless image hiding method based on the adversarial network, which can successfully resist steganalysis tools. Liu et al. [3] proposed a coverless image hiding algorithm based on DenseNet feature image retrieval and DWT sequence mapping, which has better robust and security performance resisting image attacks.

The above methods indicate that the coverless information hiding method has been well studied in the image field. Compared with images and videos, the text has less redundant information, so it is more difficult to implement text information hiding techniques [4]. However, the text has several advantages: simple

encoding, extensive data, small space occupation, and frequent use, so the text has good concealment and superior research value. At the same time, text information hiding technology has attracted the attention and interest of many researchers because of its great value in wireless transmission [5], secret communication, copyright issues, and other aspects. Besides, with the development of text information hiding and automatic text generation technology [6], text steganalysis technology is also constantly developing, which has brought a serious threat to information hiding.

As a natural carrier, text generates trillions of data every day, which is exceptionally suitable as an information hiding carrier. There are two main types of coverless text information hiding methods, the search method [7] and the generation method [8,9]. Both ways are developing rapidly, but they are limited by the development of natural language processing technology. When the length of secret information is long, the search method's implementation is complex, and the generation method may have semantic ambiguity, sentence failure, poor readability, and other problems. Moreover, both of them have insufficient embedding capacity. Zhang et al. [10] proposed to build a text database, using the word level and frequency of the secret information to match in the database, find the appropriate text to send. Although this method does not need to modify the carrier, which reduces the possibility of being attacked, the embedding rate still needs improvement. Lu et al. [11] proposed a coverless question camouflage method combined with random codes, which using secret information to generate a camouflaged form of an exam question. This method avoids the direct transmission of secret information, reduces the possibility of being discovered, and improves hidden capacity. Mo et al. [12] embed information in HTML documents by inserting invisible characters (such as spaces, tabs) in web pages, but detection resistance is poor. Synonym replacement [13,14] improves the hidden capacity and success rate, currently the best semantic-based coverless information hiding algorithm. Zhao et al. [15] proposed using high-frequency function words in Chinese to hide information, which has a low success rate of hiding. Liu et al. [16] proposed the method to extract all the components of Chinese characters and increase the capacity of information hiding by using part of speech to hide the number of keywords. Long et al. [17] proposed a text coverless information hiding based on word2vec, and it uses word2vec to obtain similar keywords. When text retrieval fails, keywords can replace similar keywords, increasing the hiding success rate and slightly increasing the hiding capacity. Long et al. [18] also proposed to use Web text to hide information, but this has significant instability. It is related to real-time web pages, and its hiding success rate is volatile. Although researchers have proposed many information hiding methods, their security and robustness are still challenging to meet actual needs.

To solve the above problems, we propose a robust text coverless information hiding method based on multi-index. The main contributions of this work can be summarized as follows:

1. We propose a multi-index secret information transmission method. In this method, a piece of secret information can generate multiple indexes. Even if a third party broke one or loses part of the carrier, the receiver can still extract the index's information. It can significantly improve the robustness.
2. The secret information can be extracted by recombination of its multiple groups. We used random increment factors to control the keyword's order, which can accurately extract secret information.
3. We use a multi-index robust method to extract multiple sets of secret information. It can judge whether the carrier is attacked by evaluating the group's continuity of secret information, which dramatically improves the security.

The following chapters of this paper are as follows: Section 2 mainly introduces the related work. Section 3 detailly introduces the proposed method and shows the secret information transfer process. The experiment results are given in section 4. Finally, section 5 provides the conclusion.

2 Related Work

2.1 HanLP and TF-IDF

Recently, image and text processing technology have developed rapidly [19–22]. How to accurately segment sentences into words has been a research hotspot in natural language processing [23]. Word segmentation is the process of recombining consecutive word sequences into word sequences according to certain specifications. In order to promote natural language processing in the production environment, HanLP is proposed, a Java toolkit composed of a series of models and algorithms. It appears the character of clear architecture, up-to-date *corpus*, complete functions, customizable, and efficient performance. HanLP's word segmentation rate can reach 20 million words per second in extreme speed mode.

After text segmentation, it is often necessary to analyze words in the text. In natural language processing, the most commonly used methods are word frequency statistics and word *TF-IDF* (word frequency-inverse document frequency) feature extraction. *TF-IDF* is a weighting technique widely used in information retrieval and text mining. As a statistical method, *TF-IDF* is used to evaluate the importance of a word to document set or document in a *corpus*. The core idea is that this word or phrase has a good classification ability and is suitable for classification if a word appears in an article with a high frequency of TF and rarely appears in other articles.

The text segmentation and the calculation of the *TF-IDF* features are essential, mainly for preparing the subsequent topic model clustering.

2.2 Text Topic Distribution

LDA (Latent Dirichlet Allocation) is a document topic generation model known as a three-layer Bayesian probability model containing a three-layer structure of documents, topics, and words. LDA has achieved great success in text topic clustering and mining by introducing hyperparameters that control model parameters to the feature word layer, text collection layer, and topic layer [24]. In recent years, scholars have begun to apply the LDA topic model to big data platforms. Spark is one of the leading big data platforms, and the distributed memory design architecture makes its running speed 5 to 150 times faster than traditional Hadoop. The Spark platform provides LDA topic model clustering methods based on online and EM implementation methods. The LDA topic clustering method of the EM method relies on the graph computing module in Spark to implement and is suitable for cluster parallel computing. Fig. 1 shows a schematic diagram of EM LDA topic clustering based on the Spark platform. The primary process is to collect the data source and extract the data on the Spark platform through text segmentation cleaning and calculation of the TF-IDF features of the words in the text, and then input the features into the LDA topic model for training and finally obtain the text topic distribution.

3 The Proposed Multi-Index Robust Approach

In this section, we will detail the proposed multi-index robust mechanism, mainly composed of six parts. 1) Robust text coverless information hiding framework based on multi-index. 2) Keyword sequence control. 3) Text search containing keywords. 4) Multi-index robust text search algorithm. 5) Robust information hiding algorithm. 6) Robust information extraction algorithm. The notation used in this paper is shown in Tab. 1.

3.1 Robust Text Coverless Information Hiding Framework Based on Multi-Index Method

Fig. 2 shows the proposed framework, which comprises six parts: 1) Codebook construction and text preprocessing. 2) Segment secret information into keywords. 3) Keyword id conversion and introduce a random increment factor to control the keyword sequence. 4) Find all texts containing the keyword id in the codebook. 5) Find multiple sets of indexes through the robust text search algorithm and send them to

the receiver. 6) The receiver disassembles the indexes through the index construction protocol and extracts the secret information by deduplication through an increment factor.

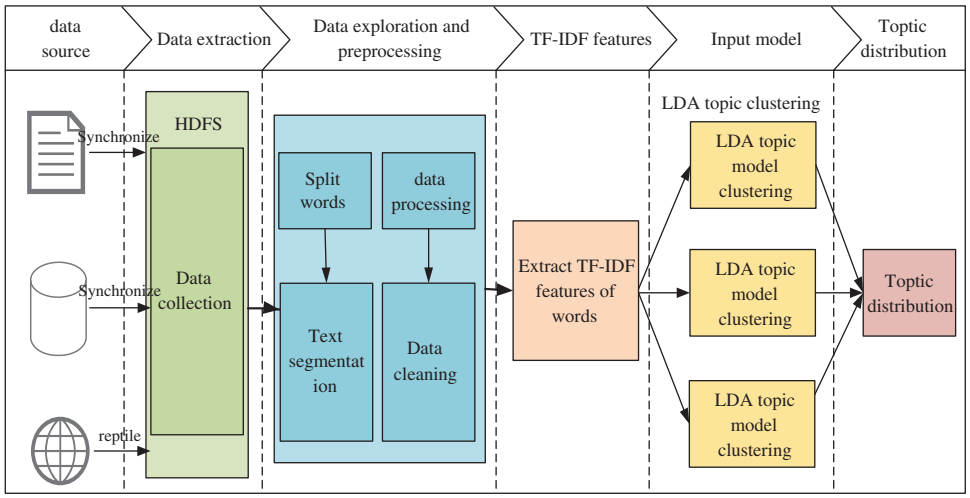

Figure 1: Schematic diagram of Spark EM LDA theme model

Table 1: The parameters

Notations	Descriptions
I	Secret information to be hidden
HanLP(I)	Perform word segmentation on *I*
keyword	Keywords after *HanLP* participle
CV(keyword)	Convert *keyword* to *id*
random	Random increament factor
id	Keyword *id*
TW(id)	Find all text containing the *id*
label	Best hidden text label
FI(label)	Convert *label* to text topic distribution
FTF(id,label)	Find word *TF-IDF* value and word frequency according to word *id* and *label*
bws{}	The best text collection containing secret information
Tws{}	Secret keyword collection contained in the ciphertext
Mts{}	All text collections containing keywords

The method requires establishing a text-topic distribution index, a global word index, and a text-word *TF-IDF* codebook. The word index comprises all the words in the text database, word frequency ranking, and the corresponding word frequency. The text index includes the topic clustering distribution of the text, and the text tag number is used to label the text containing the secret information.

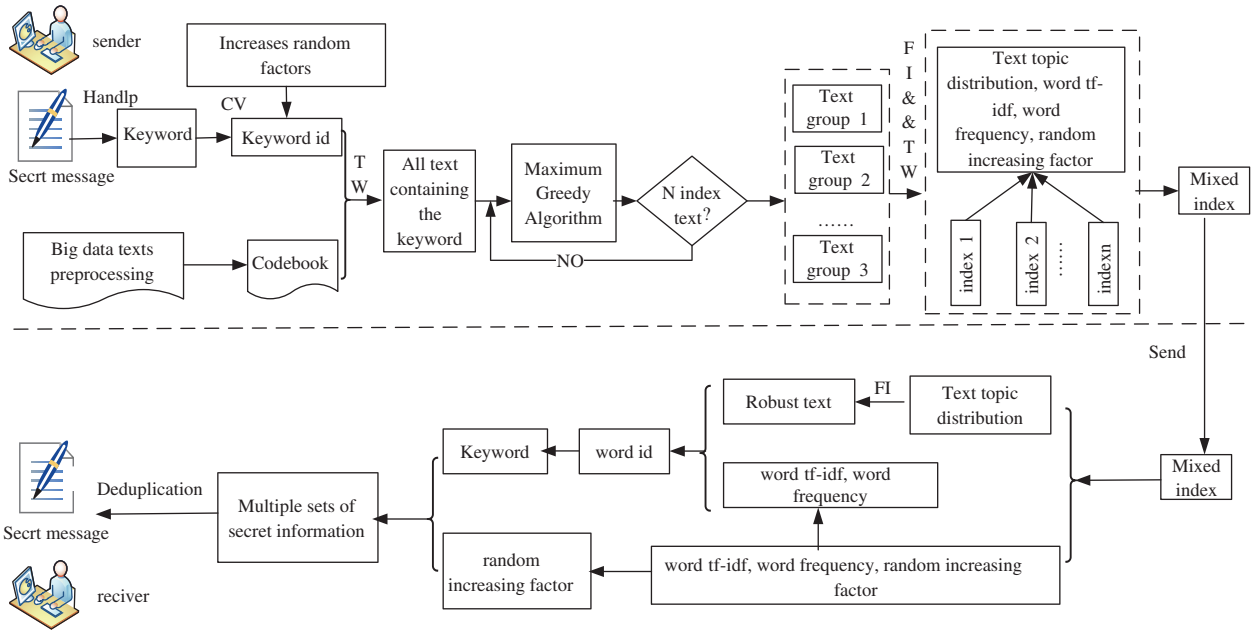

Figure 2: Robust text coverless information hiding framework

3.2 Keyword Sequence Control

In this paper, a random increasing factor is introduced to extract the secret information, and its primary functions are as follows:

1. The multi-index robust method adopted in this paper can extract multiple sets of secret information. The random increment factor can deduplicate multiple secret information disorderly.
2. Since the secret information extracted in the first step is out of order, the increment factor can reorganize the keywords in the correct order.

The example of random increment factors is shown in Fig. 3. Suppose we need to hide the secret information "文本无载体信息隐藏." First, we segment the secret information into "文本, 无, 载体, 信息, 隐藏" and the random increment factor generated are "11,34,55,65,236". We use four sets of indexes to hide secret information. Index one extracts secret information is "信息无隐藏载体," index two is "文本信息无" index three is "隐藏信息文本," index four is "载体文本." Each keyword has a random increment factor. The keywords are sorted and deduplicated by random increment factor, finally, obtain the secret information is "文本无载体信息隐藏."

Besides, double-layer random control is used for the random increment control mechanism to ensure better randomness. The behind random number must be larger than the previous random number. The algorithm as follows, initialize the random number R and take the remainder of R, then generate the corresponding random increment factor. Details see in Algorithm 1:

3.3 Text Search with Keywords

This paper improves the robustness of text coverless information hiding, which is necessary to query all texts containing secret keywords. The steps are as follows: 1) Segment secret information into keywords. 2) Keyword id conversion. 3) Find all texts containing the keyword id in the codebook.

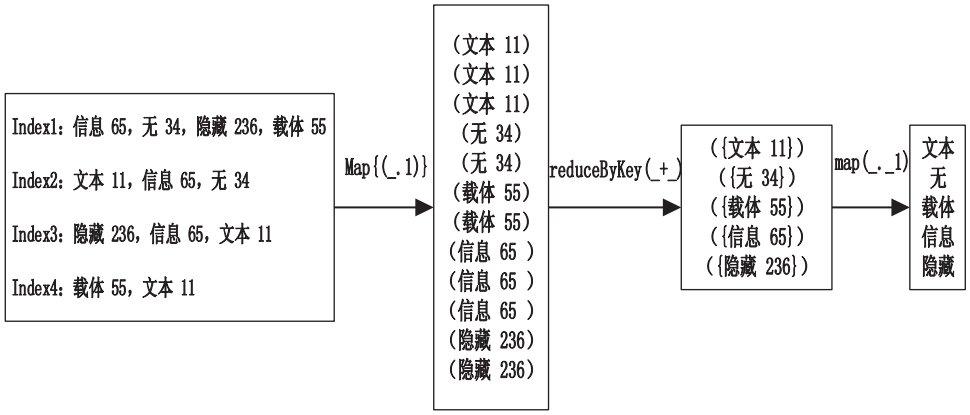

Figure 3: Example of using random increment factor

Algorithm 1: Control increment random factor

Input: initial random number *R*
Output: the random number of w_i
1: Parameter: branch number N
2: if *R* not exist:
3: Generate initial random number *R*
4: else:
5: $q = R\%N$;
6: Switch(q):
7: Case 0: generate a random number $r_0 \in [1, (q+1)*10]$
8: Case 1: generate a random number $r_1 \in [(q+1)*10, (q+2)*10]$
9: …
10: Case N-1: generate a random number $r_{N-1} \in [(q+N-1)*10, (q+N)*10]$
11: $w_i = R + r_q$
12: Return w_i

This paper uses the HanLP Chinese word segmentation tool to segment the secret information (I). As shown in Eq. (1):

$$I = \{w_1, w_2 \ldots w_k\} \tag{1}$$

$w_i (1 \leq i \leq k)$, where w represents a keyword segmented using *HanLP*.

We convert the segmented keywords into keyword id through the word index *CV*. Shown in Eq. (2):

$$w_{id-i} = CV(w_i) \tag{2}$$

Where the w_{id-i} represent the word *id*.

For example, the secret information “百度新闻是包含海量资讯的新闻服务平台” is cut into the keywords “百度”, “新闻”, “是”, “包含”, “海量”, “资讯”, “的”, “新闻”, “服务”, “平台.” The word, word frequency, and corresponding id number are shown in Tab. 2.

Algorithm 2 is the text search algorithm designed in this paper. It mainly consists of the following steps: Loop through all the words, each time it traverses all the text that contains the secret keyword, and

at the same time generates a random increment number for each word. Finally, it returns all texts containing secret keywords.

Table 2: Word, word frequency, and corresponding id

words	counts	id
百度	998	4608
新闻	8870	500
是	381487	3
包含	1953	2586
海量	437	8759
资讯	847	5295
的	1661749	0
新闻	8870	500
服务	23036	141
平台	8491	532

Algorithm 2: Find the text that contains secret information keywords

Input: $\boldsymbol{I} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k\}$
Output: all texts that contain secret information keywords: ***Mts***
1: All text collections containing keywords: *Mts* = null
2: for i = 1 to length (I) do
3: Convert to the word: $w_{id-i} = CV(w_i)$
4: Generate random for each w_i
5: Text collections containing the words: $Mts_i = TW(w_{id-i})$
6: Append to all text collection: Mts. append (Mts_i)
7: end for
8: return all texts that contain secret information keywords: Mts

3.4 Robust Multi-index Search Algorithm

The multi-index robust algorithm is implemented based on the greediest algorithm, an optimization process for all the ciphertexts found, and the best-hidden text is selected with the least amount of hidden text each time. By traversing all texts to find the text that contains the most secret keywords, a comparison is made every time the text is searched, and it is determined that the text already contains the most secret information, and the loop is traversed until all the keywords are found. This paper designs a multi-index robust algorithm as shown in Algorithm 3:

Algorithm 3: Find the best-hidden text

Input: $\boldsymbol{I} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k\}$
Output: multiple best texts contain secret messages: ***bws***
1: Return multiple index: *bws* = *{}*
2: for i ← 1 to n // n represents several sets of indexes
3: Returns the set of text with the most keywords: bts_i = null
4: While *I*! = null:
5: Current best-hidden text: *bts* = null
6: Record the keywords that have been included: *wsc* = null
7: for *text, words* in *Tws*:
8: Take the intersection: *covered* = *I* ∩ *words*
9: if *length*(*covered*) > *length*(*wcs*) :
10: *bts* = *text*
11: *wsc* = *covered*
12: Remove included words: *I* -= *wsc*
13: Add to the best text collection: bts_i*.add(bts)*
14: Add to multi-index collection: *bws.append*(bts_i)
15: Remove the best text collection found: *Tws* = *Tws.remove*(bts_i)
16: Return multiple best texts contain secret messages: *bws*

3.5 Robust Information Hiding Algorithm

After finding the multi-index text based on the above method, this section focuses on the robust information hiding algorithm. This method mainly includes the following processes:

1) According to Eq. (1), the secret information *I* will be segmented and divided into several keywords w_i.

2) The keyword w_i is converted into the corresponding keyword *id* according to Eq. (2), and a random increment factor is generated for each keyword to ensure that the receiver can extract its keywords in an orderly manner.

3) Find all text collections containing keywords, defined as *all_text*, as shown in Eq. (3)

$$all\_text = TW(word_{id}) \tag{3}$$

4) Find multiple sets of best texts according to the multi-index robust text search algorithm, defined as *best_text*. For *best_text*, the corresponding text's secret keywords in best_text can be constructed, which are recorded as *SCRET_WORDS*. Then use Algorithm 3 to find the best-hidden text *text_label*.

5) According to the text index codebook, the best *text_label* is converted into *Spare*, a multi-group text topic index distribution, as shown in Eq. (4).

$$Spare = FI(text\_label) \tag{4}$$

6) Find the corresponding word frequency and *TF-IDF* feature according to the text-word *TF-IDF* codebook, as shown in Eq. (5).

$$\left(word_{count}, word_{tf}\right) = FTF(word_{id}, text\_label) \tag{5}$$

where the $word_{tf}$ represents the keyword *TF-IDF* value, and $word_{count}$ represents word frequency.

7) The *TF-IDF* feature index included of $word_{count}$, $word_{tf}$, random increment number of each keyword, which is recorded as *IDFIndex*. Generate multiple sets of *IDFIndex* and *Spare* and send to the receiver.

3.6 Robust Information Extraction Algorithm

The receiver disassembles the indexes according to the index construction protocol and then deduplicates and sorts to extract the secret information. The steps of a secret robust information extraction algorithm are as follows:

1) **Disassemble the index**. The receiver extracts the mixed index containing multiple sets of secret information to obtain *Spare* and *IDFIndex*.

2) **Obtain the hidden text label**. Using Eq. (4), the receiver obtains the *text_label* of the hidden text based on the topic distribution index.

3) **Obtain the topic distribution.** Using Eq. (5), convert the text_label to the $word_{count}$ and $word_{tf}$.

4) **Get the keyword.** The receiver obtains the keyword id through the $word_{count}$, $word_{tf}$, then converts it into a keyword.

5) **Reorganize and extract the information**. Finally, extract the secret information through the random increment factor.

4 Experimental Results and Analysis

4.1 Experimental Environment

In the experiment, the *corpus* is mainly from the Chinese *corpus* of Sogou Lab, which is divided into six categories: social, sports, tourism, education, culture, and military. In addition, multiple sets of experimental results verify the method proposed in this article. The experiment adopts a distributed structure, so the experiment development environment is on a personal PC, completed by IntelliJ IDEA. Place the codebook on the two computing nodes of Spark and the work on the personal PC.

4.2 The Evaluation Indicator

In this paper, the experiment refers to the algorithm of the reference Long et al. [18]. Test data comes from 120 texts provided by the reference [18]. We divide these texts into 1k-6k, a total of 120, with words ranging from 1 to 2000. The text carrier comes from the Sogou Lab news data set.

1) Hidden success rate. Defined as the ratio of successfully hidden to the secret information, denoted by P_i, the definition is shown in Eq. (6). Where x represents the actual number of hidden characters and X represents the number of characters that need to hide.

$$P_i = \frac{x_i}{X_i} \ (\mathrm{i} = 1, 2, \ldots, 120) \tag{6}$$

2) Average hiding success rate. Because this paper focuses on improving robustness, random loss of 5%, 10%, 15%, 20%, 25%, 30%, 35% of carriers to test the recovery rate of secret information. Calculate the average value for each missing test. Use the average value of all P_i as the average hiding success rate, denoted by $\bar{P}_r$, as shown in the definition Eq. (7). Where P_i represents the success rate of hiding each text.

$$\bar{P}_r = \sum_{i=1}^{120} \frac{P_i}{120} \ (\mathrm{i} = 1, 2, \ldots, 120) \tag{7}$$

3) The overall hiding success rate. This paper sets seven missing cases, generates hidden success rates seven times, and calculates the average value. Use $\bar{SR}$ to represent the overall hiding success rate. The definition is shown in Eq. 8. Where $\bar{P}_{ri}$ respectively represent the average hidden success rate of 5%, 10%, 15%, 20%, 25%, 30%, and 35% of the carrier loss. Such as $\bar{P}_{r1}$ represent loss 5% carrier and $\bar{P}_{r2}$ represent loss 10% carrier

$$\bar{SR} = \frac{\sum_{r_i=1}^{n} \bar{P}_{ri}}{n} \quad (r_i = 1, 2, 3, \ldots, n) \tag{8}$$

4.3 Experimental Results

4.3.1 Robustness Comparison

To make the experiment more convincing, we calculate the average value of 120 text hiding success rates when the carrier randomly loses 5%, 10%, 15%, 20%, 30%, and 35%. We also compared this method with the reference method [18] (Long's method). The result is shown in Fig. 4.

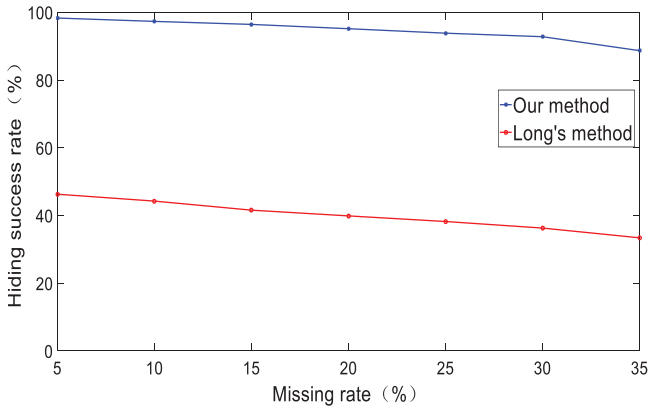

Figure 4: Comparison of hiding success rates under different deletion rates

From Fig. 4, we can see that the hiding success rate of this paper is much higher than that of the reference [18] (Long's method) under different carrier deletion rates, and it is also relatively stable. With the carrier deletion rate increase, the hiding success rate of reference [18] decreases faster, while this method basically in balance. Because we adopt a multi-index robust mechanism, a total of multiple sets of secret information texts are searched, and the best text for each search is different. Therefore, this method can search multiple text groups, so the secret information can be extracted well even if the carrier is lost, achieving strong robustness.

We also compared this method with reference [18] in the success rate of hiding each text. As shown in Fig. 5, there are 120 1k-6k texts, comparing this paper with the reference [18] in the case of different carrier deletion rates, and each text lost 5%, 10%, 15%, 20%, 25%, 30%, 35% randomly. Picture (a) shows the case where the carrier randomly loses 5%. From (a) to (g) corresponding to the carrier loss of 5% to 35%, we can see that the secret information hiding success rate of the reference [18] is not high even though the carrier deletion rate is minimal. As the carrier deletion rate increases, the success rate of information hiding is much worse. However, the success hiding rate in this paper is very stable, and it can be seen that the method in this paper has strong robustness. According to Eqs. (6)–(8), the seven times average hiding success rate in this paper has reached 94.89%.

Meanwhile each text is randomly lost 5%, 10%, 15%, 20%, 25%, 30%, 35% of the text carrier. We respectively compared the length of secret information hiding of this method and reference [18]. As shown in Fig. 6, figure (a) to (g) correspond to carrier loss of 5% to 35%, figure (a) shows the case where the carrier loses 5% randomly. We can see from the figure that the method adopted in reference [18] has inferior stability, while our method is very stable. As the carrier loss rate increases, the number of hidden words in this paper also fluctuates slightly, but overall, it has a powerful performance compared to the reference [18]. From figure (a) to (g), the lost more the carrier, the shorter the secret information

hidden length in the reference [18], while this paper maintains a stable state. We can see that the method proposed in this paper is very robust. Figure (h) is a random loss of 5%, 10%, 15%, 20%, 25%, 30%, 35% of each text, and averaged to get the comparison between the length of the secret information in each experiment and the length of the actual hidden secret information. We can see that, within a certain range of carrier loss, the number of Chinese characters successfully hidden by this method will not change greatly due to the change of the secret information length. But the reference [18] (Long's method) will change a lot. The 120 texts are 120 secret messages. Long's method represents the length of secret information that can success hide when the carrier is randomly lost using the method of reference [18]. Our method represents the length of secret information that can success hide under the condition of random carrier loss using the method proposed in this paper.

4.3.2 Security Analysis

Since this paper uses a multi-index, setting three sets of indexes can extract three sets of secret information. We can determine whether the carrier has been tampered with by comparing the three groups of extracted secret information, thereby ensuring the security of confidential information.

For example, we can extract three sets of secret information after we hide the secret information "中天杯第九届中国上海苏州河城市龙舟国际邀请赛开赛比赛共邀请境内外支龙舟队参赛." We can compare the secret information consistency to determine whether the carrier has tampered.

Fig. 7 shows three sets of secret information extracted by three sets of indexes:

We can see that the three groups of secret information extracted are entirely consistent. It can be determined that the carrier has not been modified, thereby ensuring the security of secret information.

At the same time, the secret information is converted into easy-to-express digital numbers during the information hiding. The text index includes the topic cluster distribution of the text and the text tag number used to label the text containing secret information. The transmitted indexes are all digital numbers. Even if the index is stolen during the transmission process, it is not easy to identify the stealer because the index is an abstract number. Therefore, the proposed method has strong security while ensuring robustness.

4.3.3 Secret Hiding Rate Comparison

We also tested the hiding rate of different lengths of secret information hiding. Select 1k to 6k texts (20 texts in each group) as the secret information and compared the transmission rate of this method with the reference [18], shown in Tab. 3.

It can be seen from Tab. 3 that the hiding rate of this paper is faster than that of reference [18]. Because this paper uses multiple indexes to hide multiple sets of secret information, the throughput is relatively more significant, so the hiding rate is faster. However, as the length of the secret information increases, the transmission rate of this paper is getting lower because the time to find the carrier becomes longer.

4.3.4 Analysis of Secret Information Transmission Load

In this experiment, we test the transmission load, taking the secret information 1k as an example. The transmission load required to hide 1kb secret information is shown in Tab. 4.

Where the text topic distribution represents the distribution of text topics, $word_{tf}$ represents the *TF-IDF* value corresponding to the keyword *id*, and $word_{count}$ represents the corresponding word frequency of the keyword id, and the random number controls the order of secret information keywords.

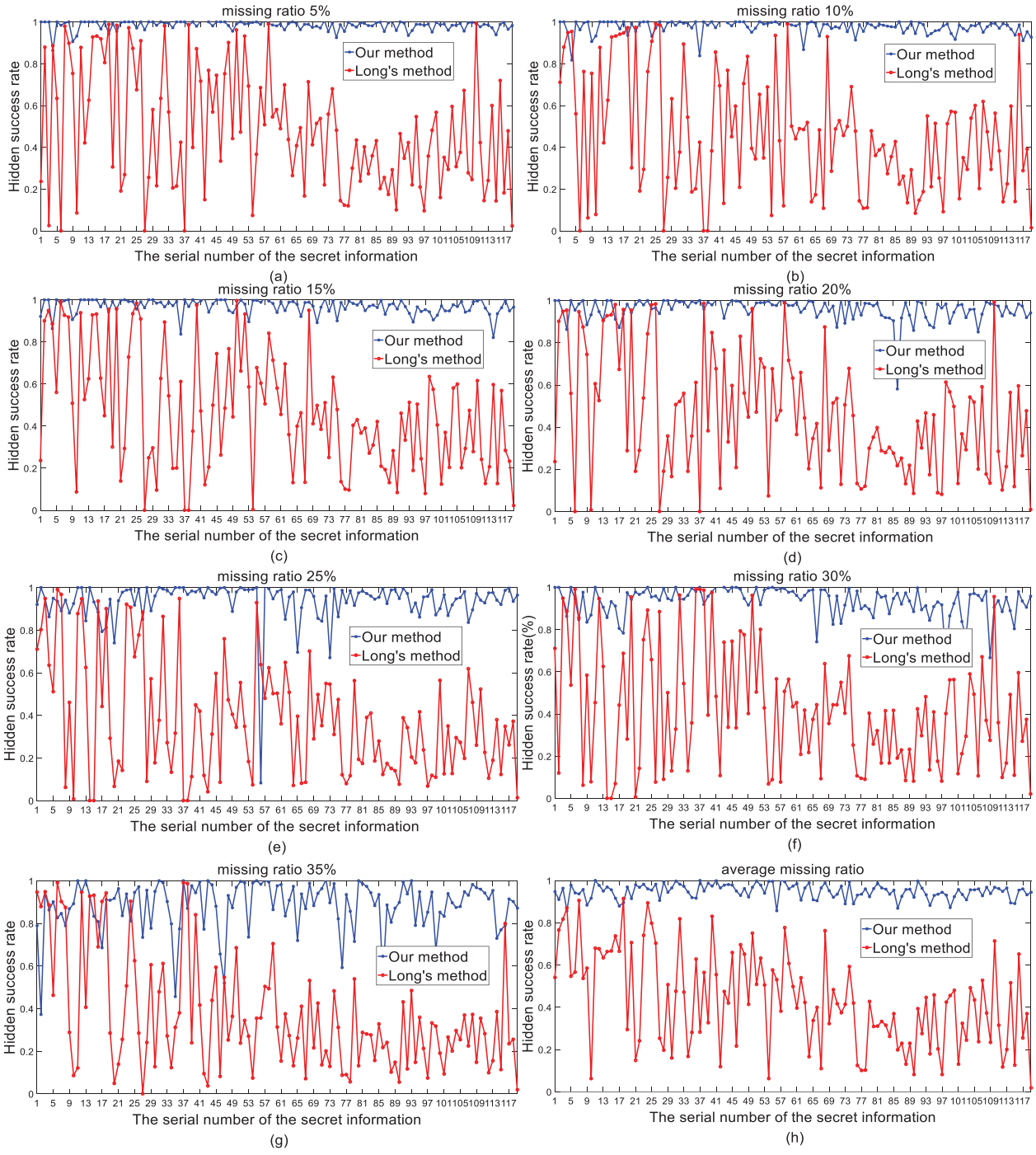

Figure 5: Experimental comparison of hidden success rate under different loss rates

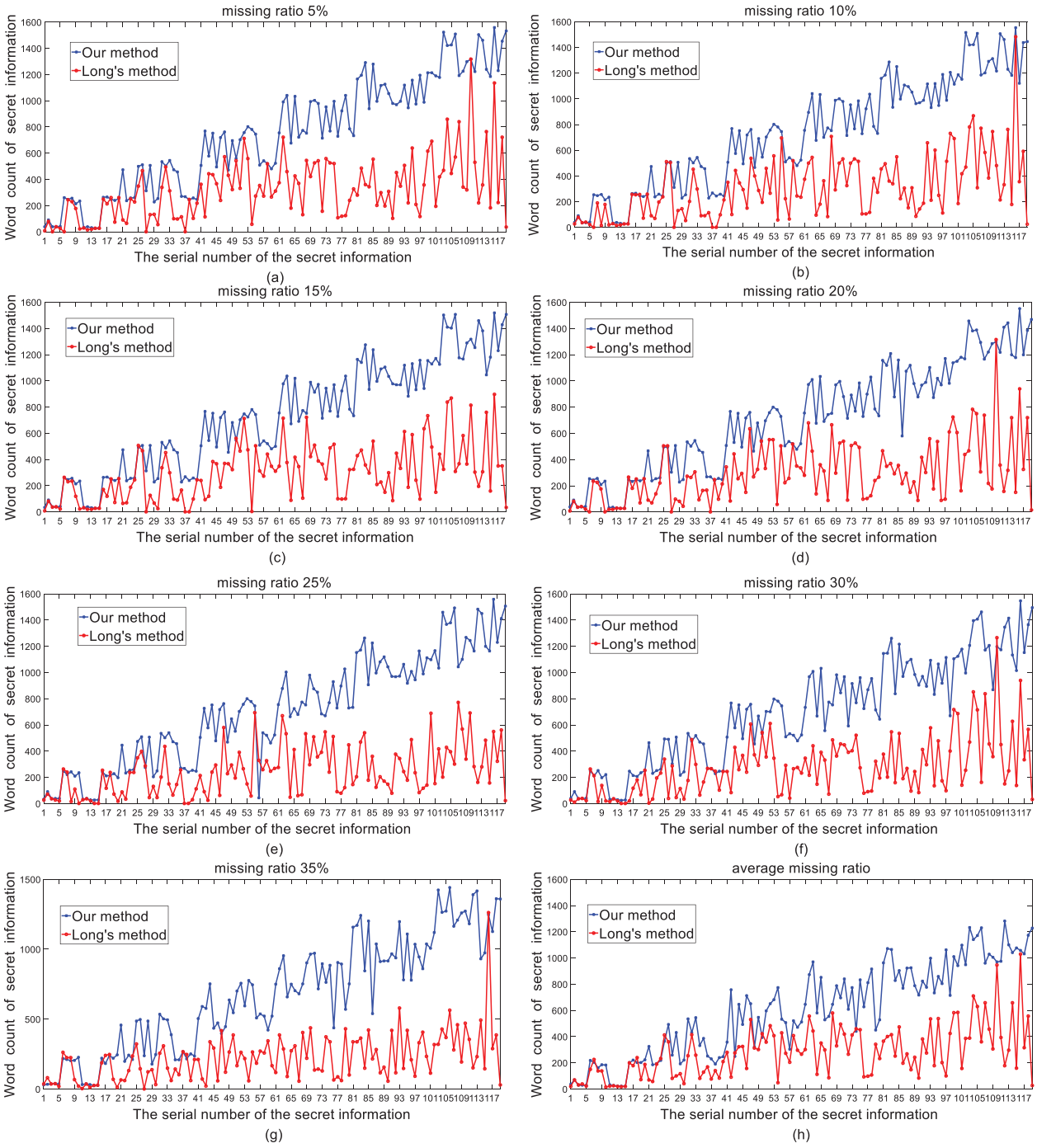

Figure 6: The number of successfully hidden Chinese characters in each experiment

所隐藏的文本label如下:
863008186
1204831969
302373396
954406953
464500320
MsgCounts1= 38
中天杯第九届中国上海苏州河城市龙舟国际邀请赛开赛比赛共邀请境内外支龙舟队参赛
所隐藏的文本label如下:
1519870359
1204831969
302373396
954406953
464500320
MsgCounts2= 38
中天杯第九届中国上海苏州河城市龙舟国际邀请赛开赛比赛共邀请境内外支龙舟队参赛
所隐藏的文本label如下:
1251952981
836203152
1462289616
491654794
464500320
MsgCounts3= 38
中天杯第九届中国上海苏州河城市龙舟国际邀请赛开赛比赛共邀请境内外支龙舟队参赛

Figure 7: Three sets of secret information extracted

Table 3: The hiding rate of different sizes of secret information hiding

Text size	1K	2K	3K	4K	5K	6K
Paper 18(bit/s)	7.79	9.51	23.59	33.64	26.02	25.58
This paper(bit/s)	57.98	54.14	40.87	35.27	33.73	32.94

Table 4: Load required to hide 1kb secret information

36 text topic distribution
249 $word_{tf}$
249 $word_{count}$
249 random number

5 Conclusion

In this paper, we proposed a robust text coverless information hiding method based on the multi-index method. In this method, the sender sends multi-indexes to the receiver to achieve better robustness. Experiments show that this paper can accurately extract secret information even if the carrier is lost when transmitting secret information, and its robustness is greatly improved. Since the original carrier has not been modified, it can resist attacks from various steganographic tools. Moreover, Multiple sets of indexes can be used to extract multiple sets of secret information. By comparing whether the secret information is

consistent to determine whether the carrier has been tampered with, the extracted secret information's quality is further improved. However, the method in this paper still has the problem that a small number of proper nouns such as person names and place names cannot be hidden, resulting in the hidden success rate not reaching 100%, which will be further optimized in the follow-up work.

Acknowledgement: The author would like to thank the support of Central South University of Forestry & Technology and the support of National Science Fund of China.

Funding Statement: This project is supported by the Degree & Postgraduate Education Reform Project of Hunan Province under Grant 2019JGYB154 and the Postgraduate Excellent teaching team Project of Hunan Province under Grant [2019]370-133, the Natural Science Foundation of Hunan Province under Grant 2020JJ4141 and the Natural Science Foundation of Hunan Province under Grant 2020JJ4140.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. J. Luo, J. H. Qin, X. Y. Xiang and Y. Tan, "Coverless image steganography based on multi-object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[2] J. H. Qin, J. Wang, Y. Tan, H. J. Huang, X. Y. Xiang *et al.,* "Coverless image steganography based on generative adversarial network," *Mathematics*, vol. 8, no. 9, pp. 1394, 2020.

[3] Q. Liu, X. Y. Xiang, J. H. Qin, Y. Tan, J. S. Tan *et al.,* "Coverless steganography based on image retrieval of DenseNet features and DWT sequence mapping," *Knowledge-Based Systems*, vol. 192, no. 2, pp. 105375–105389, 2020.

[4] I. J. Cox and M. L. Miller, "The first 50 years of electronic watermarking," *EURASIP Journal on Advances in Signal Processing*, vol. 2002, no. 2, pp. 126–132, 2002.

[5] T. Xu, M. Zhao, X. Yao and K. He, "An adjust duty cycle method for optimized congestion avoidance and reducing delay for wsns," *Computers, Materials & Continua*, vol. 65, no. 2, pp. 1605–1624, 2020.

[6] Z. Yang, S. Zhang, Y. Hu, Z. Hu and Y. Huang, "VAE-Stega: Linguistic steganography based on variational auto-encoder," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 880–895, 2021.

[7] S. W. Shi, Y. N. Qi and Y. F. Huang, "An approach to text steganography based on search in internet," in *2016 Int. Computer Sym. (ICS)*, IEEE, pp. 227–232, 2016.

[8] Z. L. Zhou, Y. Mu, N. S. Zhao, J. T. Wu and C. N. Yang, "Coverless information hiding method based on multi-keywords," in *International conference on cloud computing and security*, Cham: Springer, pp. 39–47, 2016.

[9] C. Y. Chang and S. Clark, "Linguistic steganography using automatically generated paraphrases," in *Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, pp. 591–599, 2010.

[10] J. J. Zhang, Y. C. Xie, L. C. Wang and H. J. Lin, "Coverless text information hiding method using the frequent words distance," in *Int. Conf. on Cloud Computing and Security*, Cham: Springer, pp. 121–132, 2017.

[11] L. Hai and S. L. Ping, "Carrier-free test disguise combining indirect transmission and random codebook," *Journal of Applied Science*, vol. 36, pp. 331–346, 2018.

[12] J. Mo, "Designing and implementation of HTML-text-based information hiding algorithm," *Journal of Shandong University of Technology (Natural Science Edition)*, vol. 23, pp. 21–24, 2009.

[13] C. Gan, X. M. Sun and y. L. Liu, "An improved Chinese text information hiding method based on synonym substitution," in *national academic conf. on information hiding and multimedia information security*, pp. 137–140, 2008.

[14] L. He, J. B. Lin, T. Z. Li and D. Y. Fang, "An anti-attack watermarking based on synonym substitution for Chinese text," *2009 Fifth Int. Conf. on Information Assurance and Security, IEEE*, vol. 1, pp. 356–359, 2009.

[15] M. Z. Zhao, X. M. Sun and H. Z. Xiang, "Research on natural language information hiding algorithm based on function word transformation," *Computer Engineering and Application*, vol. 042, pp. 158–160, 2006.

[16] Y. L. Liu, J. Wu and G. J. Xin, "Multi-keywords carrier-free text steganography based on part of speech tagging," in *2017 13th Int. Conf. on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, IEEE, pp. 2102–2107, 2017.

[17] Y. Long and Y. L. Liu, "Text coverless information hiding based on word2vec," *Int. Conf. on Cloud Computing and Security, Springer*, pp. 463–472, 2018.

[18] Y. Long, Y. L. Liu, Y. Q. Zhang, X. S. Ba and J. H. Qin, "Coverless information hiding method based on web text," *IEEE Access*, vol. 7, pp. 31926–31933, 2019.

[19] Z. Wang, J. H. Qin, X. Y. Xiang and Y. Tan, "A privacy-preserving and traitor tracking content-based image retrieval scheme in cloud computing," *Multimedia Systems*, 2021.

[20] W. T. Ma, J. H. Qin, X. Y. Xiang, Y. Tan and Z. B. He, "Searchable encrypted image retrieval based on multi-feature adaptive late-fusion," *Mathematics*, vol. 1019, pp. 1–15, 2020.

[21] J. H. Qin, W. Y. Pan, X. Y. Xiang, Y. Tan and G. M. Hou, "A biological image classification method based on improved CNN," *Ecological Informatics*, vol. 58, no. 4, pp. 101093, 2020.

[22] T. Q. Zhou, B. Xiao, Z. P. Cai and M. Xu, "A utility model for photo selection in mobile crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 20, no. 1, pp. 48–62, 2021.

[23] L. Xiang, S. Yang, Y. Liu, Q. Li and C. Zhu, "Novel linguistic steganography based on character-level text generation," *Mathematics*, vol. 8, no. 9, pp. 1558, 2020.

[24] Z. Zhou, J. H. Qin, X. Y. Xiang, Y. Tan and Q. Liu, "News text topic clustering optimized method based on TF-IDF algorithm on spark," *Computers, Materials & Continua*, vol. 62, no. 1, pp. 217–231, 2020.