

A Big Data Text Coverless Information Hiding Based on Topic Distribution and TF-IDF

Jiaohua Qin, Central South University of Forestry and Technology, Changsha, China

Zhuo Zhou, Central South University of Forestry and Technology, Changsha, China

Yun Tan, Central South University of Forestry and Technology, Changsha, China

Xuyu Xiang, Central South University of Forestry and Technology, Changsha, China

 <https://orcid.org/0000-0002-2778-7531>

Zhibin He, Central South University of Forestry and Technology, Changsha, China

ABSTRACT

Coverless information hiding has become a hot topic in recent years. The existing steganalysis tools are invalidated due to coverless steganography without any modification to the carrier. However, for the text coverless has relatively low hiding capacity, this paper proposed a big data text coverless information hiding method based on LDA (latent Dirichlet allocation) topic distribution and keyword TF-IDF (term frequency-inverse document frequency). Firstly, the sender and receiver build codebook, including word segmentation, word frequency and TF-IDF features, LDA topic model clustering. The sender then shreds the secret information, converts it into keyword ID through the keywords-index table, and searches the text containing the secret information keywords. Secondly, the searched text is taken as the index tag according to the topic distribution and TF-IDF features. At the same time, random numbers are introduced to control the keyword order of secret information.

KEYWORDS

Big Data Text, Coverless Information Hiding, Text Information Hiding, Text Topic Distribution, TF-IDF Features

1. INTRODUCTION

Information hiding technology, as an important branch in the field of information security, mainly uses the redundancy of human sensory organs to digital information to hide secret information in another

DOI: 10.4018/IJDCF.20210701.0a4

This article, published as an Open Access article on June 4th, 2021 in the gold Open Access journal, the International Journal of Digital Crime and Forensics (converted to gold Open Access January 1st, 2021), is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

information carrier, so that the hiding carrier information still shows the original characteristics. This information carrier can be any type of data, such as text, image, video or audio (Cox,2002, p.225). Although the external features of the hiding carrier are still retained, it still needs to change part of the information of the carrier (Zhang, 2016, p.475), which makes it unable to effectively resist replay attack, OCR technology, statistical analysis and other steganographic detection tools.

In view of the existing information hiding technology that needs to change carrier information, scholars have proposed the concept of coverless information hiding in recent years. The main idea of this method is that it does not need to modify the carrier information, and uses some specific characteristic information in the existing open carrier to hiding secret information (Zhou Z,2015, p.123). Because it does not make any modification to the carrier, it has good resistance to the detection of various steganographic tools. At present, researches on coverless information hiding mainly focus on two aspects: coverless information hiding based on image and text (Qin J,2019, p.171373). In terms of images, Zhou et al. (2016) proposed a coverless information hiding method based on image bag of word model (p.527), which used the bag of word model to extract visual keywords in each image, and constructed a mapping relation library for keywords of text information and visual keywords to hide information. Luo et al. (2020a;2020b) introduced deep learning into coverless image steganography, used semantic features and image segmentation based on Mask RCNN to hide information, which improved the robustness of the method. Liu et al. (2020) filtered images based on image retrieval of Dense Net features and used DWT (p. 105376) to generate hash sequences of images, improved the performance of steganography and expanded its application scope. Liu et al. (2018) combined with the Generative Adversarial Networks(p.371), replaced the category tag in the Generative Adversarial Networks with secret information and transmitted it as the driving generation of classified image, extracted the secret information in the classified image through discriminator in the Generative Adversarial Networks, and realized the coverless information hiding with the generation of Generative Adversarial Networks. In terms of text. Zhang et al. (2017a;2017b;2018) proposed a coverless information hiding method-based rank map. This method used word rank map and word frequency of words as distance calculation to retrieve ordinary text containing secret information from text database to realize coverless information hiding. However, this method has a low hiding capacity, and a Chinese character can only be hidden in a natural text. Chen et al. (2015) proposed coverless information hiding technology based on mathematical expressions (Sun,2002, p.707) of Chinese characters in 2015 (2015, p.133). This method first extracted the secret information vector from the secret information, and then retrieved a text containing the secret information vector based on the big data text, so as to achieve the purpose of hiding the secret information without any modification to the text. Zhou et al. (2016) proposed a coverless information hiding method based on multi-keywords to improve the capacity of hidden information (p.39). The main idea is to hide the number of keywords in the text hidden by keywords. Although this method improved the capacity of information hiding to some extent, it did not make high use of the text when indexing the text database. Liu and Wu (2017a,2017b) extracted all parts of Chinese characters, and used part of speech to hide the number of keywords to improve the capacity of information hiding. Long et al. (2018) proposed a method for text coverless information hiding based on word2vec (p.463). This method used word2vec to get similar keywords, that is, when the text retrieval fails, the similar keywords can be replaced with keywords, so that the hiding success rate can reach 100% and the hiding capacity can be slightly increased. Lu et al. (2018) proposed a coverless information hiding method combining indirect transmission and random codebook to solve the problem (p.331) that the coverless information hiding method had a small information hiding capacity and needed to build a large sample database. In the above references, although the hiding capacity has been improved, but it is still relatively small which is difficult to meet the actual demand.

Since text is the most widely used information carrier in people's daily life, especially in the background of big data era, the Internet can generate hundreds of millions of texts every day, which makes it possible to collect and integrate large amounts of texts. Therefore, text-based coverless

information hiding is a research direction with great potential. This paper proposes a mixed index method based on the text LDA topic distribution and keyword TF-IDF features in the context of big data. In this method, LDA topic clustering on the texts library through the big data platform are computed, and at the same time the TF-IDF features of the words in each text are calculated, which are constructed into codebook. When the sender sends a secret message to the receiver, the secret information is segmented into keywords at first all and the sender retrieves the secret information in the codebook. Secondly the hidden text that meets the search the conditions is merged and sent to the receiver as an index tag according to the LDA topic distribution of the corresponding text and the TF-IDF features of the keywords. As secret information is segmented and transformed, topic distribution of different texts will be different. Therefore, secret tags can effectively guarantee the security of secret information by using text topic distribution and TF-IDF feature of words as mixed index. In addition, our experiments show that the method proposed in this paper improves the hidden capacity to some extent.

The following chapters of this paper are as follows: section 2 mainly introduces some related work. Section 3 introduces the information hiding method and information extraction method in detail. Section 4 introduces the experiment and experimental results of this paper. Section 5 is the conclusion.

2. RELATED WORK

2.1 Text Segmentation and Word Frequency Features

Sentence analysis in Chinese text needs to be divided into words. How to accurately divide text into sentences has always been a research hotspot in natural language processing technology. Hanlp is an open source Java word segmentation toolkit consisting of a series of models and algorithms. It can not only provide word segmentation, but also has complete functions in lexical analysis, syntactic analysis and semantic understanding. In extreme mode, Hanlp's word segmentation rate can reach 20 million words per second.

After word segmentation, it is often necessary to analyze the words in the text. In natural language processing, word frequency statistics of words and feature extraction of words TF-IDF are the most commonly used methods. The word frequency statistical method holds that the topic words in the text often appear repeatedly in the text, so the word frequency in the text can be used as a reference for text analysis. However, there may be many meaningless function words in the text, and these meaningless words will interfere with the theme words of the text, so the desired theme words are often not obtained by simply counting the word frequency in the text. The TF-IDF method introduces the concept of word frequency-inverse text frequency, that is, only when the frequency of a certain word appears in a certain text is high, but the frequency of the word appearing in the whole text library is low, the word has a high probability to belong to the topic word. As shown in **Formula 1**, where $TF - IDF_{ij}$ represents the TF-IDF features of word i in text j , tf_{ij} represents the frequency of word i in text j , $Num(T)$ represents the number of Chinese text in the whole text library, and $Num(w_i \in D)$ represents the number of word i contained in the text library.

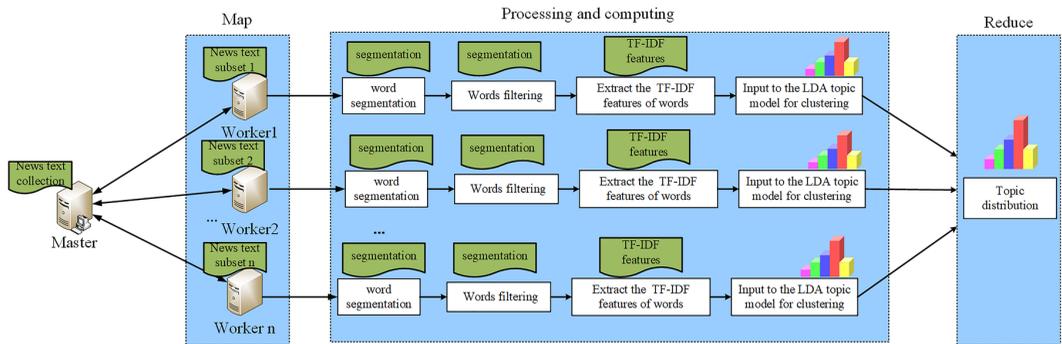
$$TF - IDF_{ij} = tf_{ij} * \log\left(\frac{N}{Num(w_i \in D)} + 1\right) \quad (1)$$

2.2 Topic Model Clustering of Big Data Text

LDA topic clustering model is a three-layer Bayesian model, which has achieved great success in text topic mining and clustering by introducing super parameters of control model parameters into text set layer, topic layer and feature word layer. With the advent of the era of big data, scholars began to

apply the LDA thematic model to the big data platform. As one of the popular big data platforms, Spark's memory-based distributed architecture is 10 to 100 times faster than a traditional Hadoop platform. Spark platform provides LDA topic model clustering method based on EM and Online. The LDA topic clustering method of EM relies on graph computing module (GraphX) in Spark, which is suitable for cluster parallel computing. **Figure 1** is a diagram of EM LDA topic clustering based on Spark platform. The main process includes the text segmentation, cleaning and calculation of TF-IDF feature on Spark platform, then the feature is input into LDA topic model for training, and finally the text topic distribution is obtained.

Figure 1. Spark EM LDA topic model



3. PROPOSED METHOD

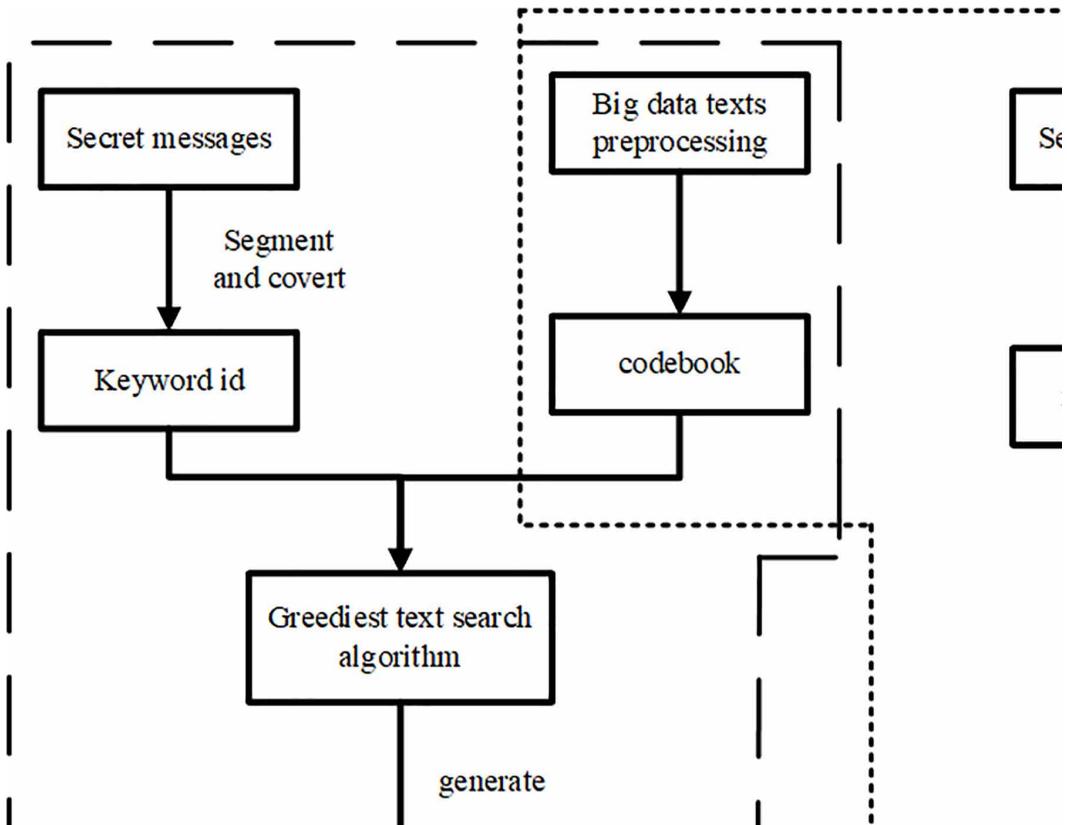
3.1 Framework of Coverless Information Hiding on Big Data Text

The main idea of coverless information hiding is to achieve the purpose of hiding secret information without modifying carrier data. Therefore, it is one of the important tasks of coverless information hiding technology to find an open data carrier containing secret information. It is difficult to find the text containing the whole secret information directly in the open data carrier. If the secret information is divided into several keywords, it will be easy to find and search through the public data carrier containing keywords. In order to query keywords in the data carrier more quickly and efficiently, it is necessary to build a data codebook that can be directly queried after processing from the open data carrier, so as to avoid the need to traverse the entire open data carrier library for each query.

This paper proposes a method of coverless text information hiding based on topic distribution and TF-IDF features mixed index. The sender and receiver use the same method to construct the codebook for the agreed public text data carrier. The sender then transmits the secret key which contains the text topic distribution and the mixed index constructed by the TF-IDF feature of the word of secret messages. The receiver uses the pre-shared key to decrypt the secret key, the index of topic distribution of hiding text and the TF-IDF index of keyword are obtained. Then the secret information will be restored through parsing index of topic distribution of hiding text and the TF-IDF index of keyword.

Figure 2 is the framework diagram of the text coverless information hiding system proposed in this paper. The system consists of four parts: big data text preprocessing, secret information segmentation and keyword id conversion according to the word index codebook, keyword id query in codebook and Greediest selection method with text contains secret to realize information hiding. The overall process is as follows: the sender and receiver construct the codebook by preprocessing the big data text, the sender, in order to ensure the security of the secret information, segments the secret information and then retrieves the text containing keywords. After the text containing keywords is obtained, the index

Figure 2. System framework



tag containing text and secret keywords is obtained through the codebook, and finally the index tag is merged and sent to the receiver to realize information hiding.

3.2 Construct Index/Codebook

The sender and the receiver must use the same method to create a codebook for the same text library before the information can be transmitted. In this paper, it is necessary to establish global word index, text-topic distribution index and text-word TF-IDF codebook. The word index is composed of all the words contained in the text library, corresponding word frequency and word frequency ranking. It is mainly used to easily express in the process of information hiding. The text index is composed of the text tag number and the topic clustering distribution of the text. Text-word TF-IDF codebook is composed of text label, word id and its corresponding TF-IDF feature in the text, which is used to locate the text and secret information.

Word index construction method is as follows:

1. Use word segmentation tools for each text in the text library word segmentation, and then calculate word frequency for all words on the Spark platform;
2. The word frequency of words is ranked in descending order, and the rank is the word id. Words, counts and id are used to build the word index codebook, where counts means the corresponding word frequency, as shown in **Figure 3**.

The text index is constructed as follows:

Figure 3. Words index

words	counts	id
的	1661749	0
在	406622	1
了	390188	2
是	381487	3
一	234677	4
和	230573	5
不	210960	6
有	188026	7
我	151178	8
也	146936	9
为	144050	10
中	129967	11
上	125743	12
人	122109	13

1. Every text in the text library is segmented by segmentation tools, and in order to ensure that each text can be uniquely represented, a hash value is generated on the text as the text label.
2. Obtain the topic clustering distribution of each text that after each word segmentation using the LDA topic model algorithm on Spark platform.
3. Each text label and its text topic distribution corresponding to constitute a text index as shown in **Figure 4**.

Text-word TF-IDF codebook construction method is as follows:

1. Calculate the word TF-IDF feature of each text and input it to the LDA topic model.
2. When the LDA model is calculated, the word id within each text and corresponding TF-IDF features will be obtained.
3. The words and corresponding TF-IDF features under each text are constructed as text - word TF-IDF codebook, as shown in **Figure 5**.

3.3 The Segmentation of Secret Information

For the convenience of description, we first explain the relevant symbols. The symbol definitions are shown in **Table 1**.

Figure 4. the text index

label	topicDistribution
437685033	[0.9989419025120423,1.5596485167842058E-4,1.4869761840358744E-4,1.2284001621429137E-4,1.421812550929398E-4,1.8195599466404205E-4,1.53516769125E-4,1.228376375 [3.606667099723375E-4,6.291421145656947E-4,0.016674277114239727,3.546385590990398E-4,4.530846284832556E-4,0.3275350258864141,6.614967972887461
621132692	[5.940233792317611E-4,8.265681687693056E-4,7.029499151700773E-4,5.794654535797946E-4,6.980153180502878E-4,0.9950779882506736,7.3299738333304271
1705083116	[2.7519104322758475E-4,3.527288634864065E-4,4.241105300587268E-4,2.6306310256556705E-4,3.5831701799964917E-4,0.9974203248947756,3.667654743979E-4,823401633 [3.1732344312519414E-4,0.08893733146408021,4.235077298287122E-4,3.163432636318471E-4,4.0564354701912695E-4,0.0698405970702743,4.31686447499489
919515702	[1.810331170594951E-4,2.3448927528228874E-4,2.0886577539596141E-4,1.771425220426673E-4,0.5124934641725023,4.4716933126543145E-4,0.486025967724E-4,775780036 [2.6506272714904295E-4,0.622269677696316,0.02515297144306988,2.6543202948094575E-4,3.695583099426292E-4,0.29983938064362553,3.886357219647194E-4,738209778 [2.8615449224190923E-4,4.216619026852731E-4,4.1181628001352047E-4,3.0631707454337566E-4,3.441119411676263E-4,7.1568844092106894E-4,0.0573829915E-4,1035785671 [0.10656626666027427,3.3344534164624794E-4,0.07983313011353418,2.0878264273734302E-4,0.5446123264927207,0.26787677178200253,2.798112149792949E-4,423866411 [5.4317392094365E-4,7.023531135770847E-4,5.941625534921722E-4,4.842827544751536E-4,6.734777537489542E-4,9.604041467777411E-4,0.827408714924474E-4

Figure 5. Text - word TF-IDF codebook

Text label	Text distribution	Word ₁	Word ₂	...	Word _i	...	Word _n
		$tf-idf_1$	$tf-idf_2$...	$tf-idf_i$...	$tf-idf_n$
		frequency ₁	frequency ₂	...	frequency _i	...	frequency _n

Table 1. symbol definitions

Notation	Definition
M	Secret Message to be hidden
T	Public text library
$WCR()$	Word-word ID conversion function (input keywords, output keyword ID,input keyword ID, output word)
$TD()$	Text index query function (input text label, output text topic distribution; Input text topic distribution, output text label)
$TW()$	Text-word TF-IDF codebook query function (input the keyword id and return all text sets containing the keyword id)
$Hanlp()$	Word segmentation tool operation
$word_{id}^d$	Keywords id
$random$	Increasing random factor
$Final_best_texts$	best texts that contain secret messages
$TEXT_WORDS$	A collection of secret keywords for each text that contain secret

In this paper, Hanlp word segmentation tool is used to segment secret information. For the whole secret information M , it is segmented into several keywords. As shown in the **Formula 2**

$$W = Hanlp(M) = \{w_1, w_2, \dots, w_k\} \tag{2}$$

where $w_i(1 \leq i \leq k)$ named as the key words.

The segmented keywords are converted into keyword ids by global word index (WCR), as shown in Formula 3:

$$w_{id-i} = WCR(w_i) \quad (3)$$

3.4 Hidden and Text Search Keywords

To query all texts containing secret keywords, and to ensure that all secret keywords can be restored by the receiver, an incremental random factor mechanism is designed in this paper to control the order of secret keyword transmission, as shown in **Algorithm 1**.

To ensure better randomness, the double-layer random control is adopted for the algorithm of incremental random control mechanism in this paper. The specific algorithm is shown in **Algorithm 2**.

3.5 The Greediest Text Search Algorithm

Greedy text search is the optimization process of all text contains secret. The idea proposed in this paper is to select the best hidden text with the minimum number of hidden texts. The algorithm is shown in **Algorithm 3**.

3.6 Merge Index

The essence of the index in the codebook is to locate the location where the secret information exists, so the index must be able to accurately provide the information hidden by the secret information. In this paper, a mixed index construction method based on LDA topic distribution and TF-IDF features is adopted. As is shown in **Figure 6**, and the specific construction method is shown below.

Step 1: determine the final text that contains secret, according to the text index codebook, convert the label containing the text that contains secret, to a text topic Distribution, which is called *Distribution*.

Step 2: determine the TF-IDF features of secret key of the text contains secret, in order to avoid the same text in the same TF - IDF features of word, an additional secret word frequency of key words in the global text library as another reference factor, so the TF- IDF features, the corresponding global word frequency and generated random number as the index of TF - IDF together, is used to retrieve the text within the words, as *TFIndex*.

Algorithm 1. Find text that contains secret information keywords

Input: $W = \{w_1, w_2, \dots, w_k\}$
Output: all texts that contains secret information keywords.
 for $i=1$ to $length(W)$ do
 $w_{id-i} = WCR(w_i)$
 Generate random for each w_i
 $Mul_text_{list_i} = TW(w_{id-i})$
 end for
 return $Mul_text_{list_1} \cup Mul_text_{list_2} \cup \dots \cup Mul_text_{list_k}$

Algorithm 2. Control increasing random factor

```

Input: initial random number R
Output: the random number of  $w_i$ 
Parameter: branch number N
if initial random number not exist:
    Generate initial random number R
else R= Result of last iteration
q= R%N;
Switch(q):
    Case 0: generate random number  $r_0 \in [1, R_0]$ 
    Case 1: generate random number  $r_1 \in [R_0 + 1, R_1]$ 
    ...
    Case N-1: generate random number  $r_{N-1} \in [R_{N-2} + 1, R_{N-1}]$ 
Return  $R + r_q$ 
    
```

Algorithm 3. Find the best hidden text

```

Input:  $W = \{w_1, w_2, \dots, w_k\}$ 
Output: best texts that contains secret messages
Final_best_texts = null //Returns the set of text with the most keywords
While W != null: // hideKeywords
    best_texts = null //Current best hidden text
    words_covered = null //Record the keywords that have been included
    for text, words in TEXT_WORDS:
        covered =  $W \cap words$  //Take the intersection
        if  $length(covered) > length(words\_covered)$ :
            best_texts = text
            words_covered = covered
    W -= words_covered //Take the difference set
    Final_best_texts.add(best_texts)
Return Final_best_texts
    
```

Step 3: construct the mixed index. Merging the *Distribution* and *TFIndex* and generating the final mixed index.

3.7 Information Hiding

Figure 7 is the specific flow chart of the information hiding method proposed in this paper.

First, use **Formula 1** to segment the secret information M into the keyword w_i .

For each keyword w_i , word index codebook retrieval is used, and **Formula 3** is used to convert its keywords into the corresponding keyword id. At the same time, in order to ensure that the keywords can be restored orderly by the receiver, increasing random factor is added to each keyword in the hiding process. That is, a random integer is generated after each query of a keyword, and a non-negative integer is added randomly on the basis of the previous random integer in the subsequent

Figure 6. Merge index

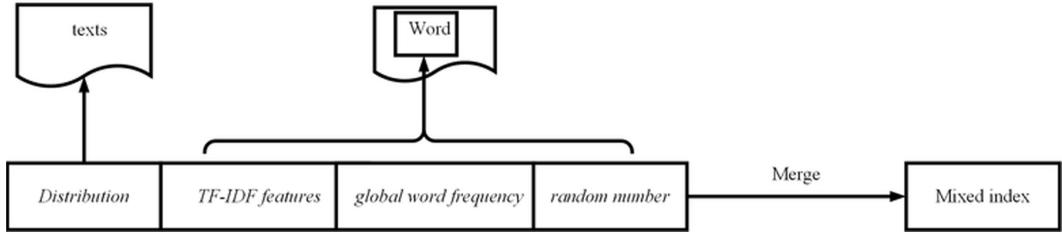
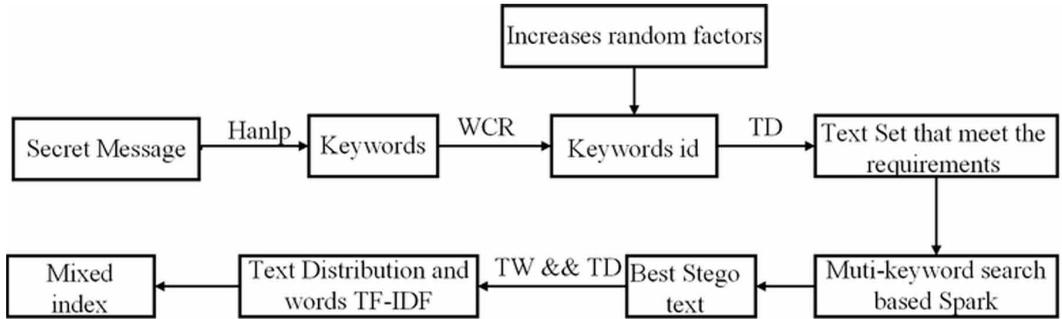


Figure 7. The process of information hiding process



query of the keyword in order to ensure the increment. The Algorithm is shown in **Algorithm 1** and **Algorithm 2**.

Find the text that contains the secret information keyword. Let the found text set containing the secret information keyword be $texts_candidate$. Where

$$texts_candidate = TD(word_{id}) \quad (4)$$

Greedy text search algorithm. Therefore, for $texts_candidate$ containing hidden keywords, you can acquire secret keyword for each text, which can be denoted as $TEXT_WORDS$. Use **Algorithm 3** on $TEXT_WORDS$ to find the label with the best hidden text.

After obtaining the best hidden text in (4), according to the text index codebook, convert the best hidden text label into the text topic distribution index $Distribution$, i.e.

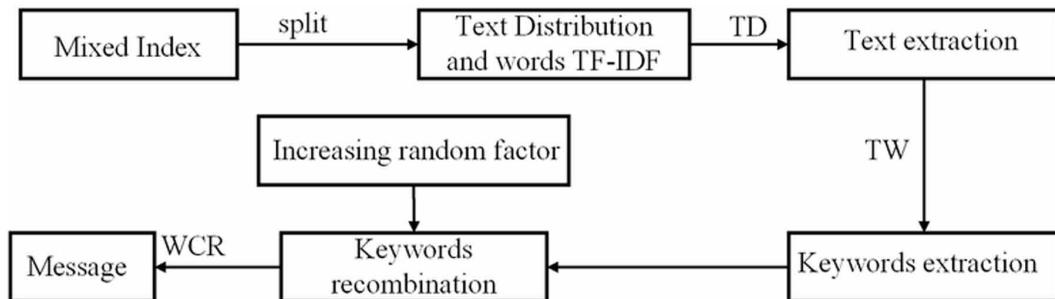
$$Distribution = TD(text_label) \quad (5)$$

The secret keyword id in the best hidden text set is searched to the corresponding TF-IDF feature and global word frequency according to the text-word TF-IDF codebook, i.e.

$$(word_{tf}, word_{count}) = TW(word_{id}) \quad (6)$$

$word_{tf}, word_{count}$ and random numbers corresponding to each keyword together constitute $TFIndex$. Finally, $Distribution$ and $TFIndex$ are Merged and sent to the receiver.

Figure 8. The process of information extraction



3.8 Information Extraction

The sender sends the mixed index to the receiver to achieve the purpose of transmitting secret information, and the receiver only needs to decrypt the mixed index and split the secret information according to the index construction protocol to restore the secret information. The steps are shown in **Figure 8**.

1. Split the index, the receiver extracts the mixed index and obtains *Distribution* and *TFIndex*.
2. Gets the hidden text, and gets the label of the hidden text in the text index codebook based on the topic distribution index.
3. Get the keyword id in the obtained text according to $word_{idf}, word_{count}$ in the *TFIndex*, and get the keyword id in the text-word TF-IDF codebook.
4. Information recombination and restoration: since random factors are generated with hidden keywords, and the overall random factors are monotonically increasing, the information can be reorganized by sorting the random factors of keyword id extracted from step 3 in ascending order. The keyword id is then restored to text information based on the word index codebook.

3.9 Security Analysis

Since the coverless information hiding method in this paper is based on big data texts, the number of carrier texts is sufficient to ensure that this method can be robust even if some carrier texts are missing. In addition, we use increasing random factor to control the sequence of secret information fragments. Even if the same secret information is hidden, the random sequence generated each time will be different, so the security of secret index is guaranteed to some extent.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental Environment

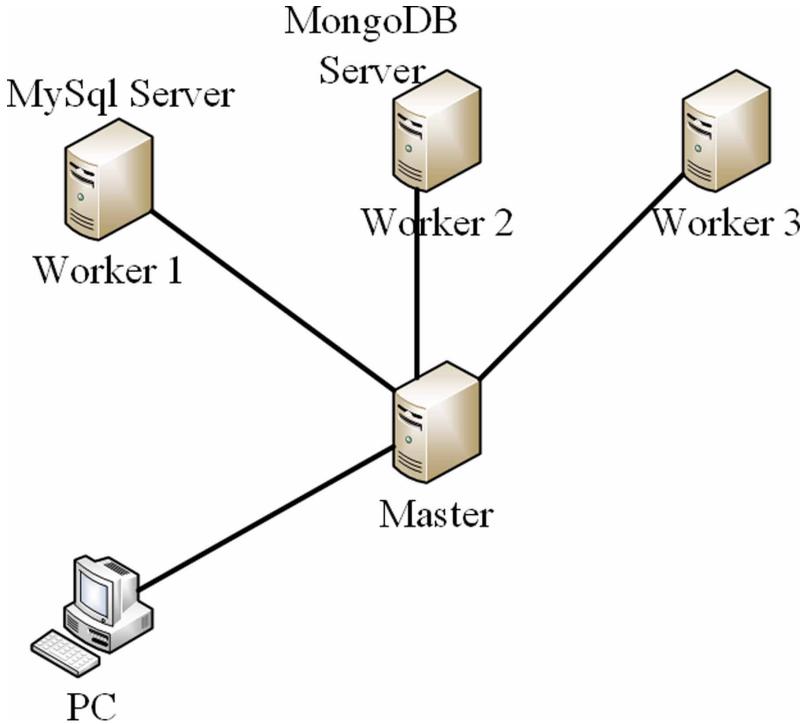
Four dawning high-performance computers in Central south university of forestry and technology based on Spark distributed architecture were used in this paper. The system, software and hardware configurations are shown in **Table 2**.

Due to the distributed structure of the experiment, the experimental development environment was completed on personal PC with Eclipse. Codebooks are placed on two of Spark's compute nodes, and work on a PC can be submitted directly to the Spark cluster via a local area network. The schematic diagram is shown in **Figure 9**.

Table 2. Experimental hardware and software configuration information

Machine	CPU	Memory	Operating System	Spark version
Master node(1 set)	8 cores	14GB	CentOS7.2	Spark2.3.0
Compute node(3set)	8 cores	8GB	CentOS7.2	Spark2.3.0

Figure 9. Experimental environment



4.2 The Evaluation Indicator

In this experiment, we refer to and reimplement its algorithm in Long (Long,2019, p.31926) and 120 Chinese texts are selected which are classified according to the length from 1KB to 6KB. The text carrier was the Sogou laboratory news data set. The hidden capacity is defined in (Chen,2017,p.313) if the Number of keywords to be hidden is k , and the Number of texts needed to hide secret information is Number, then the hidden capacity is

$$V_i = \frac{k}{Number} (i = 1, 2, \dots, 120) \quad (7)$$

After 50 experiments, the average value of all V_i was used as the average hiding capacity.

$$\bar{v} = \sum_{i=1}^{120} \frac{V_i}{120} \quad (8)$$

The success rate of information hiding is another indicator of information hiding performance, which is defined as follows.

$$P_i = \frac{x_i}{X_i} (i = 1, 2, \dots, 120) \quad (9)$$

where X_i represents the number of Chinese characters needed to be hidden in the experiment, and x_i represents the number of Chinese characters actually hidden. In the 50 experiments in the previous section, the success rate of each experiment was calculated. Similar to average hiding capacity, this paper defines the formula of average hiding capacity as follows.

$$\bar{P} = \sum_{i=1}^{120} \frac{P_i}{120} \quad (10)$$

4.3 Analysis of Experimental Results

Figure 10 is the number of secret messages and hidden successful characters each experience, we find that the success rate is not significantly affected by the length of the secret message. Figure 11 is the Hiding success rate comparison with Long (Long,2019, p.31926). According to Formula 10 and 11, the average hiding success rate in (Long,2019, p.31926) is 94.8%, but the hiding success rate of this paper is 98.24%. Figure 12 is the hiding capacity corresponding to each experiment. According to Formula 8 and 9, the average hiding capacity of this paper reaches 64.36, however, Long (Long,2019, p.31926) is 20.74.

In Figure 12, we find an extremely high point in this paper, which is due to the fact that the hidden secret information can find an identical text in the text library. If we remove this highest point, as is show in Figure 13, the average hidden capacity still remains 60.40, which is still better than Long(Long,2019, p.31926).

5. CONCLUSION

This paper proposes a coverless information hiding method based on LDA topic distribution and TF-IDF feature mixed index of big data text. This method is based on the big data text in the Internet. The sender can hide the secret information by transferring the topic model distribution of the text and the TF-IDF features of the words in the text to the receiver as mixed index . Since the method does not modify the original text carrier, it can resist the attack of various steganographic tools. In addition, this method uses massive text data as the carrier, which has stronger concealment. The method sends the mixed feature index with higher security. This method is based on parallel processing of big data, and secret information hiding adopts a greedy strategy, which to some extent increases the capacity of secret information hiding. However, this method has a small number of proprietary words cannot be hidden, resulting in the success rate of hiding cannot reach 100%. Further research will be carried out in the following work.

ACKNOWLEDGMENT

This work is supported by the Science Research Projects of Hunan Provincial Education Department (No.18C0262,18A174,19B584), the National Natural Science Foundation of China(No.61772561),the Key Research & Development Plan of Hunan Province (No. 2019SK2022),the Degree & Postgraduate

Figure 10. the number of secret messages and hidden successful characters each experience

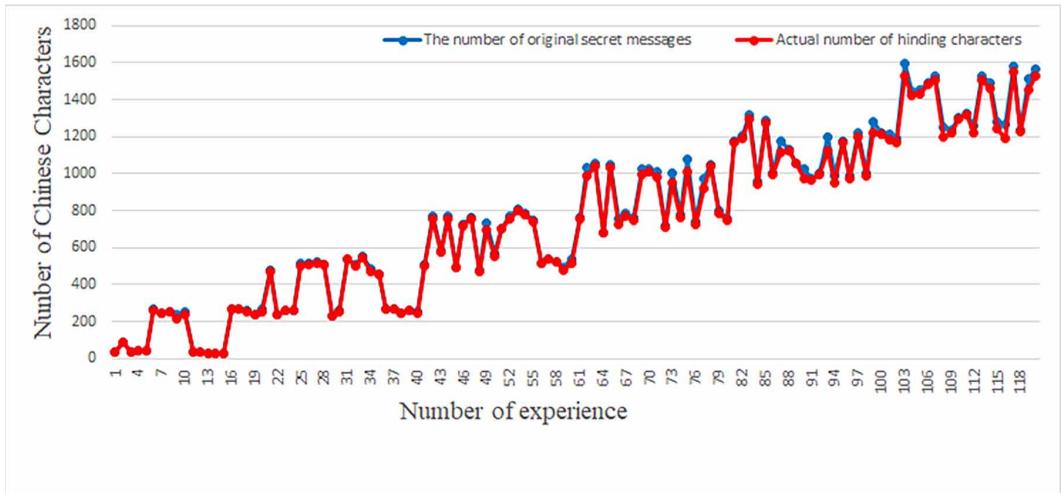


Figure 11. Comparison of the hiding success rate with Long (Long,2019, p.31926)

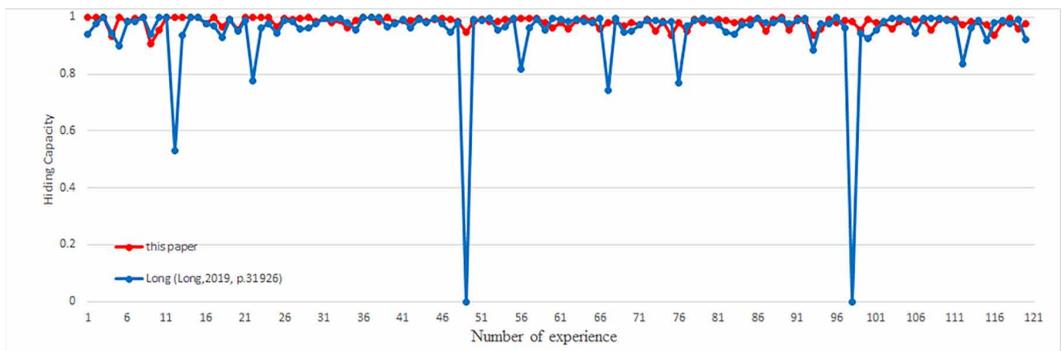


Figure 12. Comparison of hiding Capacity with Long (Long, 2019, p.31926)

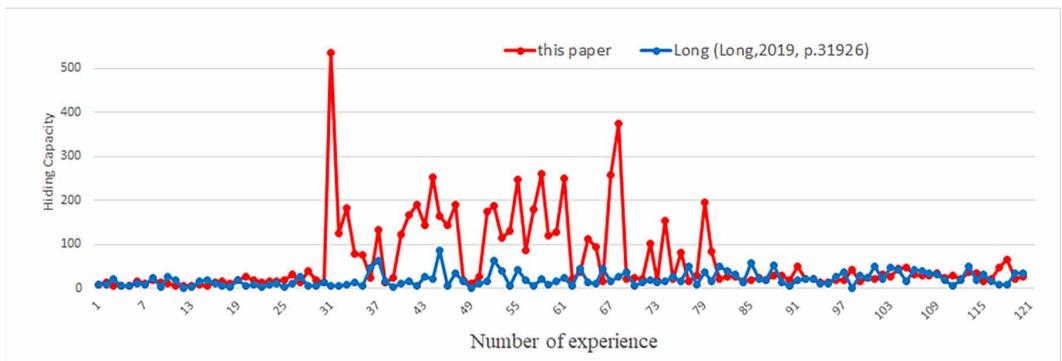
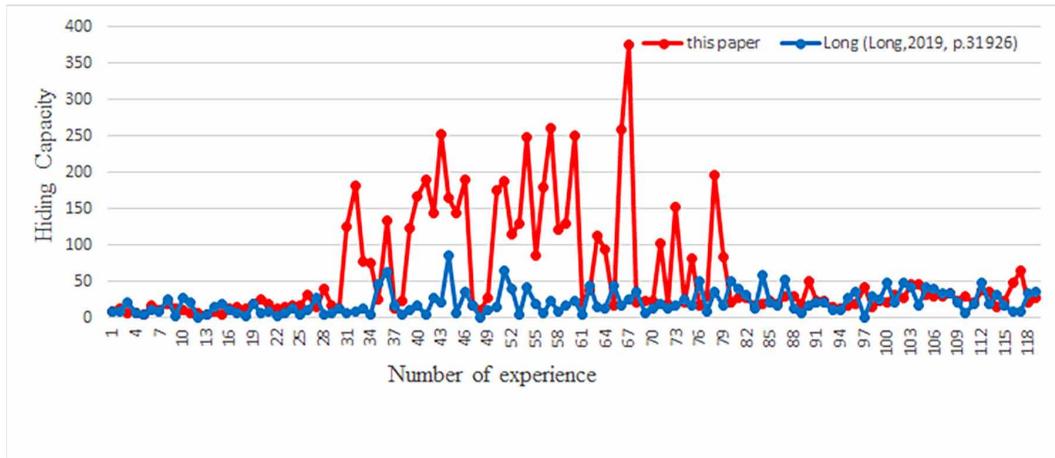


Figure 13. Comparison of hiding capacity with Long (Long,2019, p.31926) after remove highest point



Education Reform Project of Hunan Province(No.2019JGYB154),the Postgraduate Excellent teaching team Project of Hunan Province(No. [2019]370-133), the Natural Science Foundation of Hunan Province(No.2020JJ4140, 2020JJ4141),and the Postgraduate Education and Teaching Reform Project of Central South University of Forestry & Technology(No. 2019JG013).

REFERENCES

- Chen, X., Chen, S., & Wu, Y. (2017). Coverless Information Hiding Method Based on the Chinese Character Encoding. *Journal of Internet Technology*, 18(2), 313–320. doi:10.6138/JIT.2017.18.2.20160815
- Chen, X., Sun, H., & Tobe, Y. (2015). Coverless Information Hiding Method Based on the Chinese Mathematical Expression. *International Conference on Cloud Computing and Security*, 9483, 133–143. doi:10.1007/978-3-319-27051-7_12
- Cox, I. J., & Miller, M. L. (2002). Electronic watermarking: the first 50 years. *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on. IEEE*, 225–230. doi:10.1109/MMSP.2001.962738
- Liu, M., Zhang, M., & Liu, J. (2018). Coverless Information Hiding Based on Generative adversarial networks. *Journal of Applied Sciences*, 36(2), 371–382. doi:10.3969/j.issn.0255-8297.2018.02.015
- Liu, Q., Xiang, X. X., Qin, J. H., Tan, Y., Tan, J., & Luo, Y. (2020). Coverless steganography based on image retrieval of Dense Net features and DWT sequence mapping. *Knowledge-Based Systems*, 192 (2020), 105375–105389. doi:10.1016/j.knosys.2019.105375
- Liu, Y., Wu, J., & Xin, G. (2017a). Multi-keywords coverless text steganography based on part of speech tagging. *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2102–2107. doi:10.1109/FSKD.2017.8393096
- Long, Y., & Liu, Y. L. (2018). Text Coverless Information Hiding Based on Word2vec. *International Conference on Cloud Computing and Security*, 463–472. doi:10.1007/978-3-030-00015-8_40
- Long, Y., Liu, Y. L., Zhang, Y. Q., Ba, X., & Qin, J. (2019). Coverless Information Hiding Method Based on Web Text. *IEEE Access: Practical Innovations, Open Solutions*, 7, 31926–31933. doi:10.1109/ACCESS.2019.2901260
- Lu, H., & Shao, L. (2018). Combination of indirect transmission and random codebook coverless test camouflage. *Journal of applied sciences*, 36(2), 331–346. doi:10.3969/j.issn.0255-8297.2018.02.012
- Luo, Y. J., Qin, J. H., & Xiang, X. X. (2020b). Coverless Image Steganography Based on Image Segmentation. *CMC-Computers. Materials & Continua*, 64(2), 1281–1295. doi:10.32604/cmc.2020.010867
- Luo, Y. J., Qin, J. H., Xiang, X. X., Tan, Y., Liu, Q., & Xiang, L. (2020a). Coverless real-time image information hiding based on image block matching and dense convolutional network. *Journal of Real-Time Image Processing*, 17(1), 125–135. doi:10.1007/s11554-019-00917-3
- Qin, J. H., Luo, Y. J., Xiang, X. X., Tan, Y., & Huang, H. (2019). Coverless Image Steganography: A Survey. *IEEE Access*, 7(1), 171372–171394. doi:10.1109/ACCESS.2019.2955452
- Sun, X. M., Yin, J. P., & Chen, H. W. (2002). Study on mathematical expressions of Chinese characters. *Jisuanji Yanjiu Yu Fazhan*, 39(6), 707–711. doi:10.1007/978-3-319-48671-0_4
- Wu, J. (2017b). *Research on coverless information hiding method based on multi-keywords*. Hunan University.
- Zhang, J. J. (2018). *Research on coverless information hiding technology based on common words in text set*. Hunan University.
- Zhang, J. J., Wang, L. C., & Lin, H. J. (2017a). Coverless Text Information Hiding Method Based on the Rank Map. *Journal of Internet Technology*, 18(2), 127–434. doi:10.6138/JIT.2017.18.2.20160624b
- Zhang, J. J., Xie, Y., Wang, L., & Lin, H. (2017b). Coverless Text Information Hiding Method Using the Frequent Words Distance. *Lecture Notes in Computer Science*, 10602, 121–132. doi:10.1007/978-3-319-68505-2_11
- Zhang, X. P., Qian, Z. X., & Li, C. (2016). Prospect of information hiding research. *The Journal of Applied Science*, 34(5), 475–489. doi:10.3969/j.issn.0255-8297.2016.05.001
- Zhou, Z. L., Cao, Y., & Sun, X. M. (2016). Based on image coverless Bag - of - Words model of information hiding. *The Journal of Applied Science*, 34(5), 527–536. doi:10.3969/j.issn.0255-8297.2016.05.005
- Zhou, Z. L., Sun, H., & Harit, R. (2015). Coverless Image Steganography Without Embedding. In *International Conference on Cloud Computing & Security*. Springer. doi:10.1007/978-3-319-27051-7_11

Zhou, Z. M. Y., & Zhao, N. (2016). Coverless information hiding method based on multi-keywords. *International Conference on Cloud Computing and Security. Nanjing, China*, 39-47. doi:10.1007/978-3-319-48671-0_4

Jiaohua Qin received the B.S. degree in mathematics from the Hunan University of Science and Technology, China, in 1996, the M.S. degree in computer science and technology from the National University of Defense Technology, China, in 2001, and the Ph.D. degree in computing science from Hunan University, China, in 2009. She was a Visiting Professor with the University of Alabama, Tuscaloosa, AL, USA, from 2016 to 2017. She is currently a Professor with the College of Computer Science and Information Technology, Central South University of Forestry and Technology, China. Her research interests include network and information security, machine learning and image processing.

Zhuo Zhou received his BS in Network Engineering from Hunan College of humanities, Science and Technology China, in 2017. He is currently pursuing her MS in Computer Technology at College of Computer Science and Information Technology, Central South University of Forestry and Technology, China. His research focuses on Big Data and information security.

Yun Tan received the M.S. and Ph.D. degrees both from Beijing University of Posts and Telecommunications, China, in 2004 and 2016, respectively. Now she is a lecturer with College of Computer Science and Information Technology, Central South University of Forestry and Technology, China. Her research interests mainly include image security, compressive sensing and signal processing.

Xuyu Xiang received his B.S. in mathematics from Hunan Normal University, China, in 1996, M.S. degree in computer science and technology from National University of Defense Technology, China, in 2003, and PhD in computing science from Hunan University, China, in 2010. He is a professor with the College of Computer Science and Information Technology, Central South University of Forestry and Technology, China. His research interests include network and information security, image processing and machine learning.

Zhibin He received his BS in the Information Science and Technology Institute of Hunan Agricultural University, China, in 2019. He is currently pursuing his MS in information and communication engineering at College of Computer Science and Information Technology, Central South University of Forestry and Technology, China. His research interests include image processing and pattern recognition.