Coverless Text Information Hiding Method Using the Frequent Words Hash

Jianjun Zhang[1,3], Huajun Huang[2], Lucai Wang[3], Haijun Lin[3], Deng Gao[4]
(Corresponding author: Jianjun Zhang)

College of Computer Science and Electronic Engineering, Hunan University[1]
36 Lushan Rd, Yuelu Qu, Changsha Shi, Hunan 410080, China
(Email: jianjun998@163.com)
College of Computer and Information Engineering, Central South University of Forestry and Technology[2]
College of Engineering and Design, Hunan Normal University[3]
College of Software, Hunan Vocational College of Science and Technology[4]

(Received Nov. 22, 2016; revised and accepted Feb. 20 & Mar. 31, 2017)

Abstract

The attackers may discover the existence of the secret information or even get it by analyzing the cover's statistical characteristics, changes of which often occur due to the embedding. In this paper, a novel coverless text information hiding method was proposed. By using the words rank map and the frequent words hash, normal texts containing the secret information could be retrieved from the text database, and will be sent to the receiver without any modification. Because the embedding is not needed, the proposed method could be able to escape from almost all state-of-the-art steganalysis methods.

Keywords: Big Data; Coverless Information Hiding; Frequent Words Hash; Rank Map; Steganography

1 Introduction

Steganography, also known as information hiding, is a secure communication method that conveys secret messages in the form of plaintexts so that the appearances of the secret messages will not draw eavesdroppers' attention while they are being transmitted through an open channel [17]. It can be used for intellectual property protection and secret communications [9]. For example, reference [20, 25] introduced two methods of detecting illegal copies of copyrighted images. For information hiding, there are many kinds of covers, such as texts [15], images [7, 18], videos [16], etc. [1, 12, 21].

Compared to the image or other covers, text information hiding is the most difficult kind of steganography due to the lack of redundancy. Because the text is frequently used in people's daily lives, however, text information hiding has attracted many researchers' interest, and has many results [23]. Classified by the covers, text steganography could be put into three types: text format-based [3, 10, 11], generating-based and embedding-based natural language information hiding. For text format-based information hiding, the embedded information will no longer exist if the document is generated without format after extracting the text content. Generating-based natural language information hiding methods can fool the computer statistical analysis, but is relatively easy to be identified by people [2]. Embedding-based natural language information hiding methods have more robust and better concealment than text format-based information hiding, but the hiding algorithm is difficult to implement, and there are some deviations and distortions in the statistic and linguistics because of the limitation of the natural language processing [13].

Once the information hiding algorithm is public, the steganalysis methods will be appearing. So attackers will know the existence of the secret information by analyzing changes of statistical characteristics of the covers caused by the embedded information. Be there an algorithm with which the secret information could be hidden without any modification of the covers. Coverless information hiding [24, 6], firstly proposed by Xingming Sun et al., is the best answer to the above question. Reference [24] presented a coverless image steganography framework, and Reference [6] proposed a coverless text information hiding method. These two methods can directly retrieve the stego-image (stego-text) without any modification of the covers. Recently, coverless information hiding, which requires no modification on covers and could resist various steganalysis technologies, draws more and more attention from researchers [5, 22].

In this paper, a novel coverless text information hiding method is proposed. Firstly, a text database is constructed by collecting a large number of texts from the Internet. Then the word rank maps of the words will be calculated by statistically analyzing the text big data, meanwhile, the frequent words distance of every text is calculated. When a certain information will be transmitted, a normal text containing the secret information is retrieved from the text database by using the frequent words distance and the word rank maps, and sent to the receiver without any modification.

2 Coverless Information Hiding

Coverless information hiding is a new challenging research field. In fact, "coverless" is not to say that there is no carrier, but compared with the conventional information hiding, coverless information hiding requires no other carries [6]. The idea of coverless information hiding is often used in our daily life, and the acrostic poem is a classic example. An acrostic poem is shown in Figure 1 form which we can learn that the secret information is "TREE". Coverless information hiding is essentially the disclosure of secret information in the text. Its distinctive characteristic is "no embedding", that is, a carrier cannot embed secret information by modifying it [6].

Figure 1: An acrostic poem

3 The Proposed Method

3.1 Preparation of the Text Database

We construct a natural text database by fetching the news from the normal news web sites. For each word of the vocabulary, we calculate the frequency of its occurrence, and then rank the words with the descending way (Most frequent word has rank 1, next frequent word has rank 2 ...). Figure 2 shows the ranking result of words in a text database. In order to make good use of the information of the words' occurrence in a text database (or in a text), the Word Rank Map of a text database (or a text) is defined as:

$$RM = \{(w_i, f_i)|i = 1, 2, \cdots, U\} \quad (1)$$

where U is the number of unique words in a text database (or in a text), i is the rank of a word w_i, and f_i is the frequency of w_i. Figure 2 shows the word rank map of a text database. For the example in Figure 2, we can obtain

$$RM = \{(the, 124020), (and, 55654), (of, 54550), (to, 52331), \cdots\} \quad (2)$$

Obviously, the top frequent words are: the, and, of, to, in, a, on, for, and etc.

total number of words: 1967696
total number of unique words: 57618

rank	word	:	frequency:
1	the	:	124020 :
2	and	:	55654 :
3	of	:	54550 :
4	to	:	52331 :
5	in	:	46181 :
6	a	:	41390 :
7	on	:	19241 :
8	for	:	17883 :
9	that	:	17338 :
10	said	:	17230 :
11	is	:	16193 :
12	china	:	13951 :
13	with	:	13542 :
14	as	:	11784 :
15	by	:	10478 :
16	at	:	10236 :
17	it	:	9960 :
18	will	:	9435 :
19	he	:	9199 :
20	from	:	9079 :
21	has	:	8960 :
22	was	:	8861 :
23	s	:	7703 :
24	be	:	7389 :
25	have	:	7251 :
26	are	:	7145 :
27	an	:	7011 :
28	chinese	:	6992 :
29	its	:	6119 :
30	his	:	5587 :

Figure 2: Part of a text database word rank map

For each text in a text database, we can obtain its word rank map defined as Equation (1). Figure 3 shows the rank map of a text named as "2.1 million Audi cars affected by emissions cheating scandal.txt", in which there are 185 words, 113 unique words. From the rank map, we learn that the top frequent words are: the, in, software, emission, cars, etc.

For the top frequent words in the text database, we can calculate their occurrences in a text in the same collection. So, the Frequent Words Hash Function is defined as:

$$H_k(t) = \{h_1 h_2 h_3 \cdots h_k\} \quad (3)$$

where k is the number of the top frequent words chose form the vocabulary of a text database, t is a text in the text database, and h_i is defined by:

$$h_i = \begin{cases} 1 & \text{the i-th frequent word appears in text t} \\ 0 & \text{the i-th frequent word does not appear in text t} \end{cases} \quad (4)$$

total number of words: 185
total number of unique words: 113

rank	word	frequency
1	the	19
2	in	6
3	software	4
4	emission	4
5	cars	4
6	car	4
7	of	4
8	on	4
9	a	4
10	audi	3
11	to	3
12	epa	3
13	cheating	3
14	million	3
15	by	3
16	said	2
17	1	2
18	total	2
19	that	2
20	were	2
21	agency	2
22	protection	2
23	emissions	2
24	environmental	2
25	scandal	2
26	vehicles	2
27	s	2
28	u	2
29	volkswagen	2
30	company	2

Figure 3: Part of a text's word rank map

For the top 30 frequent words shown in Figure 2, we can calculate the hash value of a text named as "2.1 million Audi cars affected by emissions cheating scandal.txt". The hash value is:

$$H_k(t) = \{111111101110111101001010010010\} \quad (5)$$

So, we map a text into a 30 bits string. Figure 4 shows the hash values of some texts in a text database. In order to measure the occurrence of the frequent words in a text t, we define the Frequent Words Distance of a text as:

$$\begin{aligned} DFW_k(t) &= HD(H_k(t), (b_1, b_2, \cdots, b_k)) \\ b_i &= 0, \quad i = 1, 2, \cdots, k. \end{aligned} \quad (6)$$

where k is the number of the top frequent words, t is a text in the text database, and HD is Hamming Distance calculating operation.

```
1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 0 1 0 1 0 0 0 1 0 1 1 0
1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 0 1 1 1 1 0 1 0 1 0 0 1 0 0 1
1 1 0 1 1 1 1 1 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 0 1
1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 0 1 1 0 1 0 1 0 0 1 0 0 1
1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 0 1 1
1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 0 1 0
1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 0 1 0 1 1 1 0 0 1 1 1 0 0 0
1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0
1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 0 0 1 0
1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 0 1 1 1 0 1 1 1 0 0 1 0
1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 0 1 1 1 1 1 0 1 0 0 0 0 0 0
1 0 1 1 1 1 1 0 1 1 0 0 0 0 1 1 1 1 0 0 1 0 0 1 0 0 1 0 1 0
1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 0 1 1
1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 0 1 0 1 1 0 1 0
1 1 1 1 1 1 1 1 0 1 1 0 1 0 1 0 1 1 0 0 0 0 0 0 1 0 1 0 1 0
1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 0 0 1 1 0 0 0 1 1 1 1 1 0
1 1 1 1 1 1 1 1 1 0 1 0 1 0 1 1 1 1 1 1 0 1 0 0 1 0 1 0 0 1
1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 0 0 1 1 0 1 1 1 1 0 1 0
1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 0 0 0 0 1 1 1 0 0 0 1 1 0 1 0
1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 1 0 1 0 1 0 1 0 1 1 0
1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 1 0 0 1 1 1 1 1 0 1 0
1 1 1 1 1 1 1 1 0 0 0 0 1 0 1 1 1 0 0 0 0 1 1 0 1 0 0 0 0 1
1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 0 0 0 0 1
1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1
1 1 1 1 1 1 1 0 0 0 1 0 1 0 1 1 0 1 0 1 1 0 0 1 0 1 0 0 0 1
1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1
```

Figure 4: The hash values of some texts in a text database

By statistical analyzing the occurrence of each word in the text database, we can calculate the word rank map of each word. For each word w_i appearing in the text database, its word rank map is defined as

$$\begin{aligned} RMW_i &= \{(rw_{ij}, fw_{ij}, wt_{ij}) | i = 1, 2, \cdots, U; \\ & \qquad j = 1, 2, \cdots, N\} \end{aligned} \quad (7)$$

where rw_{ij} is the rank of the word w_i in a text wt_{ij} according to its occurrence, fw_{ij} is the frequency of w_i's occurrence in wt_{ij}, and N is the number of the texts in which w_i appears. Figure 5 shows the word rank map of "from", whose rank is 20 in the word rank map as shown in the Figure 2.

Word-'from'

Rank--Frequency--Source Text

Rank	Frequency	Source Text
46	3	'Containing China' a Japanese strateg
22	4	'Cyber Monday' sales set to hit recor
20	3	'Disturbance' only harms China, Japan Fu Yin
42	2	'Hobbit' vanished far earlier than researcher:
52	1	'In' leads by 7 percentage points ahead of Br
6	7	'Maternity tourism' in US sees chang
13	5	'Mini Messi', 5, aims to follow in star's foo
15	4	'No more survivors' in collapsed overpas
21	2	'Plane wreckage' found in Thailand fuels talk
52	3	'Slave' women held for 30 yrs rescue
41	2	'Symbol of rebirth' opens in NY
151	1	1 killed, 51 injured in Mexico candy factory
20	3	1 killed, dozens wounded at UN base in S. Sud:
19	2	1 survived, 53 bodies recovered in Algeria pl:
16	3	10 Afghan militants killed in army operatio
13	4	10 facts about HIVAIDS on World AIDS Da
12	2	10 killed in flash flood in eastern Indi
133	1	10 US Navy sailors detained by Iran repor
113	1	100 killed in tribal clashes in Sudan's Darfu
39	2	11 trapped miners rescued from S. African min
30	2	12 China-made helicopters delivered to Cambod
44	2	12 IS militants killed in air strike in weste
60	2	125 killed in month-long battles of Iraq's Anl
18	4	126 hostages rescued, 4 attackers killed as B
6	4	126 rescued, 4 attackers killed as operation :
27	1	12th Elephant Festival held in Nepa

Figure 5: Part of the rank map of "from"

3.2 Information Hiding

The information hiding process is shown in Figure 6. Detail procedures are introduced as follows. Suppose the constructed text database is T, and the communication key is k. We can calculate the word rank map of T by using Equation (1), and get the vocabulary of T ,and let it be $W = \{w_i | i = 1, 2, \cdots, U\}$ where i is the rank of w_i, and U is the number of unique words in T. For each text t_i in T, we can calculate its word rank map by using Equation (1), and let it be RM_{t_i}.

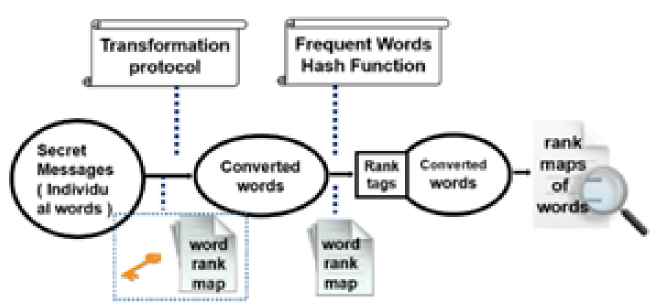


Figure 6: The process of information hiding

Because the key is k, we arrange the top frequent $k \times k$ words in W as the right part of Figure 7. Suppose the hidden message is $M = m_1, m_2, \cdots, m_n$ where m_i is a word, and n is the number of words in the hidden message. For each word m_i in M, it is chosen from the top frequent $k \times k$ words in W. Obviously, the selection range of m_i depends on k. Therefore, both sides of communication can choose more k so that m_i has more options.

w_1 ⟸	w_1	w_2	w_3	$\cdots$	w_k
w_2 ⟸	w_{k+1}	w_{k+2}	w_{k+3}	$\cdots$	w_{2k}
$\vdots$ ⟸	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
w_{k-1} ⟸	$w_{(k-2)\times k+1}$	$w_{(k-2)\times k+2}$	$w_{(k-2)\times k+3}$	$\cdots$	$w_{(k-1)\times k}$
w_k ⟸	$w_{(k-1)\times k+1}$	$w_{(k-1)\times k+2}$	$w_{(k-1)\times k+3}$	$\cdots$	$w_{k\times k}$

Figure 7: The words conversion table

3.2.1 Words Conversion

In order to enhance the security of the secret message, we convert each word in M into one of the top frequent k words in W before the information hiding. The conversion rule is shown in Figure 7. For each word in $\{w_1, w_2, \cdots, w_k\}$, it will be converted into w_1. For each word in $\{w_{(k+1)}, w_{(k+2)}, \cdots, w_k\}$, it will be converted into w_2. And so on.

For each word m_i in M, we can get its rank by using the word rank map of T, and let it be R_{m_i}. Then, it is located in the $((R_{m_i}-1)/k+1)$ row, $((R_{m_i}-1)\%k+1)$ column in the word conversion table shown in Figure 7, where "%" is a remainder operation. Therefore it will be converted into $m'_i = w_{((R_{m_i}-1)/k+1)}$. In this way, we can convert the secret information $M = m_1, m_2, \cdots, m_n$ into $M' = m'_1, m'_2, \cdots, m'_n$, and M' is a subset of the k top frequent words in W.

3.2.2 Searching the Stego-text

For each word $m'_i = w_{((R_{m_i}-1)/k+1)}$ in M', the stego-text is get as follows:

Firstly, for each text t in the text database, we calculate the hash value of $H_k(t)$, defined as Equation (3), where k is the communication key. Then, we can get the frequent words distance of t by using Equation (6), ant let it be $DFW_k(t)$.

Secondly, because m'_i is in the top k frequent words in the word rank map of T, we can get the rank map of m'_i by using Equation (7), and let it be $RMW_{m'_i}$.

Thirdly, by using the word rank map $RMW_{m'_i}$, we retrieve all texts containing m'_i to search a text t in which the rank of m'_i is equal to $DFW_k(t)$, and the frequency of m'_i's occurrence is equal to $((R_{m_i}-1)\%k+1)$. There may be some texts satisfying this condition, then, we can select a text from those texts as the stego-text for m'_i.

Finally, as described above, we can search a stego-text set for each m'_i in M'. These stego-texts is a normal text set that contains the converted secret message, and they can be sent to the receiver without any modifying.

3.3 Information Extraction

The process of extraction is shown in Figure 8. Suppose the stego-text is S, so S is a set of normal texts. The number of texts in S is the number of words in secret message M. Let k be the communication key. Because the text database T is open for all users, receiver can calculate the word rank map of T, and get the top k frequent words in W by using the communication key k. Certainly, receiver can get the same word conversion table shown in Figure 7. For each stego-text t in S, the details of information extraction will be introduced as follows.

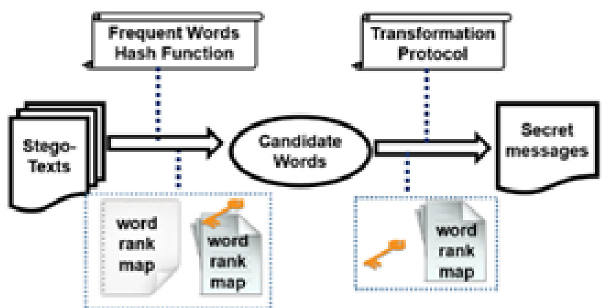


Figure 8: Secret message extracting process

3.3.1 Get the Candidate Word

Receiver can calculate the word rank map of t by using Equation (1), and the frequent words distance of t by using Equation (6) and let it be $DFW_k(t)$. By retrieving the word rank map of t, receiver can get the candidate word whose rank is equal to $DFW_k(t)$ in text t. Obviously, the candidate word is $m'_i = w_{((R_{m_i}-1)/k+1)}$.

3.3.2 Get the Secret Message

By using the word rank map of t, receiver can find the word frequency of m'_i in t, and let it be Fm'_i. Obviously, Fm'_i is equal to $((R_{m_i}-1)\%k+1)$. Receiver can find the secret message m_i that is located in the "$w_{((R_{m_i}-1)/k+1)}$" row, the $((R_{m_i}-1)\%k+1)$ column in the word conversion table shown in Figure 7.

So, receiver can get every word m_i in M. and then get the secret message $M = m_1, m_2, \cdots, m_n$.

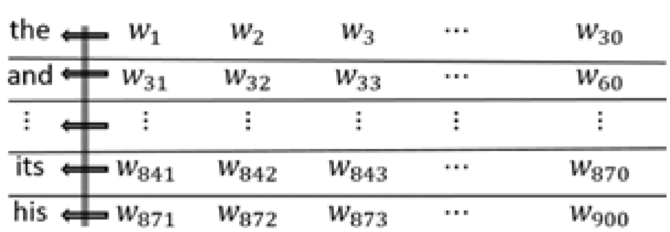
4 Discussion

4.1 An Example

In order to clearly explain the above coverless text information hiding process, we illustrate it by a simple example. We have constructed a text database which can be expanded constantly, and it is open for all users. Suppose the communication key is 30 and the secret information M is "mutual visit". It is worth mentioning that, however, M is a subset of the top 900 frequent words in the text database. Both sides of communication may choose larger k, so that words of M have more options. The operating procedure of information hiding is introduced as follows:

Firstly, sender computes the rank map of the text database and ones of each text in it. Because the communication key is 30, sender can obtain the top 30 frequent words set $W_{top30} = \{w_i | i = 1, 2, \cdots, 30\}$, and the W_{top30} is: $\{the, and, of, to, in, a, on, for, that, said, is, China, with, as, by, at, it, will, he, form, has, was, s, be, have, are, an, Chinese, its, his\}$.

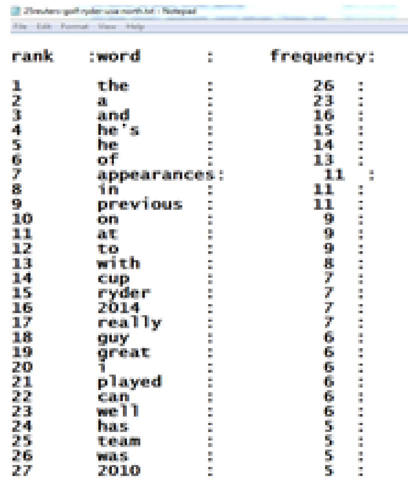
So, sender can get the word conversion table shown in Figure 9. By retrieving the word rank map of text database, sender finds the rank of "mutual" is 605, and the rank of "visit" is 183. According to the word conversion table shown in Figure 9, therefore, "mutual" is located in 21st row, 5th column, and "visit" is located in 7th row, 3rd column. Hence, "mutual" will be converted into "has", and "visit" will be converted into "on".



the	w_1	w_2	w_3	$\cdots$	w_{30}
and	w_{31}	w_{32}	w_{33}	$\cdots$	w_{60}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
its	w_{841}	w_{842}	w_{843}	$\cdots$	w_{870}
his	w_{871}	w_{872}	w_{873}	$\cdots$	w_{900}

Figure 9: The word conversion table when key is 30

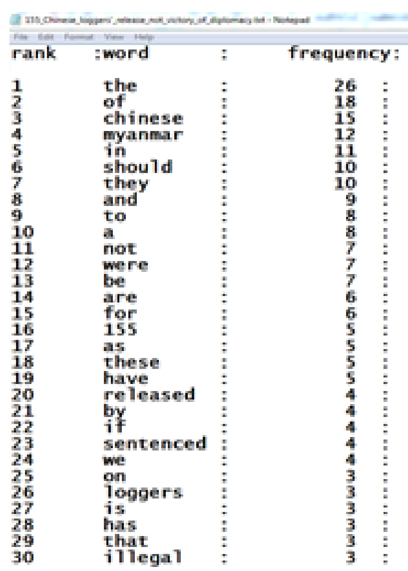
Secondly, sender calculates the word rank map of "has" and ones of "on". By retrieving the two word rank maps, sender can find a text named "25reuters-golf-ryder-usa-north.txt", and let it be t_1, and a text named "155_Chinese_loggers'_release_not_victory_of_diplomacy.txt", and let it be t_2. Their word rank maps are shown in Figure 10 and Figure 11. From Figure 10, we learn that the rank of "has" is 24 and its frequency is 5 in text t_1 whose frequent words distance is 24, and the frequent words are $\{the, a, and, of, in, on, at, to, with, has, was, is, that, as, his, for, be, an, it, by, will, he, s, have\}$ in it. From Figure 11, we learn that the rank of "on" is 25 and its frequency is 3 in text t_2 whose frequent words distance is 25, and the frequent words are $\{the, and, of, to, in, a, on, for, that, is, China, with, as, by, it, will, he, from, has, be, have, are, an, Chinese, its\}$ in it.



rank	:word	:	frequency:
1	the	:	26 :
2	a	:	23 :
3	and	:	16 :
4	he's	:	15 :
5	he	:	14 :
6	of	:	13 :
7	appearances	:	11 :
8	in	:	11 :
9	previous	:	11 :
10	on	:	9 :
11	at	:	9 :
12	to	:	9 :
13	with	:	8 :
14	cup	:	7 :
15	ryder	:	7 :
16	2014	:	7 :
17	really	:	7 :
18	guy	:	6 :
19	great	:	6 :
20	i	:	6 :
21	played	:	6 :
22	can	:	6 :
23	well	:	6 :
24	has	:	5 :
25	team	:	5 :
26	was	:	5 :
27	2010	:	5 :

Figure 10: The word rank map of a stego-text

Finally, sender sends the two texts t_1, t_2 as the stego-texts to the receiver.



rank	:word	:	frequency:
1	the	:	26 :
2	of	:	18 :
3	chinese	:	15 :
4	myanmar	:	12 :
5	in	:	11 :
6	should	:	10 :
7	they	:	10 :
8	and	:	9 :
9	to	:	8 :
10	a	:	8 :
11	not	:	7 :
12	were	:	7 :
13	be	:	7 :
14	are	:	6 :
15	for	:	6 :
16	155	:	5 :
17	as	:	5 :
18	these	:	5 :
19	have	:	5 :
20	released	:	4 :
21	by	:	4 :
22	if	:	4 :
23	sentenced	:	4 :
24	we	:	4 :
25	on	:	3 :
26	loggers	:	3 :
27	is	:	3 :
28	has	:	3 :
29	that	:	3 :
30	illegal	:	3 :

Figure 11: The word rank map of a stego-text

Because the text database is open to all users, receiver can calculate its word rank map, the top 30 frequent words and the word conversion table shown in Figure 9 by using the communication key $k = 30$. Then, he (or she) calculates the frequent words distance of t_1 and ones of t_2, and finds that they are 24 and 25. So, receiver retrieves the word rank maps of t_1 and t_2, and gets the candidate words "has" and "on" whose ranks are 24 and

25. From the word rank map of the stego-texts, he (or she) also learns that the candidate words' frequency are 5 and 3. Finally, receiver gets the secret message "mutual" located in "has" row, 5th column in the word conversion table shown in Figure 9, and "visit" located in "on" row, 3rd column. Hence, receiver successfully extract the secret message "mutual visit" from the stego-texts.

4.2 Security Analysis

Steganalysis is usually performed through the use of irrelevance between the embedded information and the carriers. Attackers often make steganalysis by analyzing the difference of their statistical distributions [19]. In our proposed hiding method, however, the carriers are normal pure text and the secret information is not been embedded in the carriers. The carriers can be sent to receiver without any modification. So the information hiding does not change the probability distribution of the carriers. According to the definition of the security of an information hiding system in [4], the proposed information hiding method is theoretically safe. At the same time, the proposed approach is also followed the Kerckhoffs Principle [8] in cryptography, and detail of information hiding is open. If he does not know the communication key, the attacker cannot gain any information about the hidden information [14]. Therefore, the proposed method could resist almost all kinds of current steganalysis method.

4.3 The Importance of Big Data

However, it is worth mentioning that, in order to enhance security, there are two works must be done: one is to change periodically the communication key to ensure that the secret message may be converted into different subsets of the top frequent words in the text database. The second is to establish a large text database (text big data) to increase the probability of the frequent words distance is equal to the rank of a word in a text, and so there are more choices of the stego-texts [23]. For example, in the chose text database, for the word "China", there are 3258 texts in which it appears, and their frequent words distance are shown in Figure 12. From the Figure 12, we learn that these values are not evenly distributed. There are 37 texts whose frequent words distance is 12, and there is only one text in which the rank of "China" is 12. Therefore, the text big data is necessary to ensure the smoothly implement of the proposed method.

Because the text big data is an important guarantee of the smooth implementation of the proposed method, some files should reside in the memory buffer when the big data is handling. We firstly calculate the word rank map of each text in the text big data, then the word rank map of the text database, and finally the ones of each word of the vocabulary, so the computing cost is expensive especially when computing the word rank map of the text database. In order to reduce the complexity, we will use the "inverted index" for storage optimization.

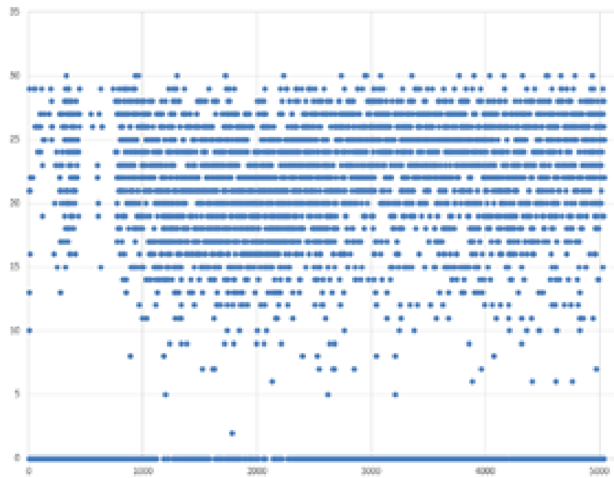


Figure 12: The distribution of frequent words distances of texts containing "China"

Because the location lab is simply designed, the capacity of the proposed method is one word per text. In order to increase the capacity of information hiding, we will design better lab location methods in the future.

5 Conclusion

This paper presented a coverless text information hiding method based on the frequent words hash. By using the words rank map and the frequent words hash, normal texts containing the secret information could be retrieved from the text database, and will be sent to the receiver without any modification. Because there is no embedding, the information hiding does not change the probability distribution of the covers. Therefore, the proposed method is theoretically safe, and could be able to escape from almost all state-of-the-art steganalysis methods.

Acknowledgments

This work is supported by National Natural Science Foundation of China (61304208), Open Fund of Demonstration Base of Internet Application Innovative Open Platform of Department of Education (KJRP1402), Open Fund of China-USA Computer Science Research Center (KJR16239), Open Fund of China-USA Computer Science Research Center (KJR16239), Hunan Province Science And Technology Plan Project Fund (2012GK3120), Scientific Research Fund of Hunan Province Education Department (13CY003, 14B106), Changsha City Science and Technology Plan Program (K1501013-11), Hunan Normal University University-Industry Cooperation Fund, and Youth Scientific Research Foundation of Central South University of Forestry & Technology (QJ2012009A)

References

[1] E. O. Blass, T. Mayberry, G. Noubir, and K. Onarlioglu, "Toward robust hidden volumes using write-only oblivious RAM," in *ACM SIGSAC Conference on Computer and Communications Security (CCS'14)*, pp. 203–214, Scottsdale, USA, 2014.

[2] S. Bo, Z. Hu, L. Wu, and H. Zhou, *Steganography of Telecommunication Information*, Beijing: National Defense University Press, 2005.

[3] J. T. Brassil, S. H. Low, and N. F. Maxemchuk, "Copyright protection for the electronic distribution of text documents," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1181–1196, 1999.

[4] C. Cachin, "An information-theoretic model for steganography," in *The Second Workshop on Information Hiding*, pp. 306–318, Oregon, USA, 1998.

[5] X. Chen, S. Chen, and Y. Wu, "Coverless information hiding method based on the chinese character encoding," *Journal of Internet Technology*, vol. 18, no. 2, pp. 91–98, 2017.

[6] X. Chen, H. Sun, Y. Tobe, Z. Zhou, and X. Sun, "Coverless information hiding method based on the chinese mathematical express," in *The First International Conference on Cloud Computing and Security (ICCCS'15)*, pp. 133–143, Nanjing, China, 2015.

[7] L. Huang, L. Tseng, and M. Hwang, "The study on data hiding in medical images," *International Journal of Network Security*, vol. 14, no. 6, pp. 301–309, 2012.

[8] S. Katzenbeisser, F. Petitcolas, *Information Hiding Techniques for Steganography and Digital Watermarking*, Artech House Publishers, 2000.

[9] T. Y. Liu, W. H. Tsai, "A new steganographic method for data hiding in microsoft word documents by a change tracking technique," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 1, pp. 24–30, 2007.

[10] S. H. Low, N. F. Maxemchuk, and J. T. Brassil, "Document marking and identification using both line and word shifting," in *IEEE International Conference on Computer Communications (Infocom'95)*, pp. 853–860, Boston, USA, 1995.

[11] S. H. Low, N. F. Maxemchuk, and A. M. Lapone, "Document identification for copyright protection using centroid detection," *IEEE Transactions on Communications*, vol. 46, no. 3, pp. 372–383, 1998.

[12] T. Mayberry, E. O. Blass, and A. H. Chan, "Efficient private file retrieval by combining ORAM and PIR," in *The Twentieth Annual Network & Distributed System Security Symposium*, pp. 1–11, San Diego, USA, 2013.

[13] P. Meng, L. Huang, Z. Chen, W. Yang, and M. Yang, "Analysis and detection of translation based steganography," *ACTA Electronica Sinica*, vol. 38, no. 8, pp. 1748–1852, 2012.

[14] F. A. Petitcolas, R. Anderson, and M. Kuhn, "Information hiding - A survey," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1062–1078, 1999.

[15] L. Y. Por, T. F. Ang, and B. Delina, "WhiteSteg: A new scheme in information hiding using text steganography," *WSEAS Transactions on Computers*, vol. 7, no. 6, pp. 735–745, 2008.

[16] S. Wang, C. Xiao, and Y. Lin, "A high bitrate information hiding algorithm for video in video," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 3, no. 11, pp. 2572–2577, 2009.

[17] N. Wu, M. Hwang, "Data hiding: Current status and key issues," *International Journal of Network Security*, vol. 4, no. 1, pp. 1–9, Jan. 2007.

[18] Z. Xia, X. Wang, X. Sun, Q. Liu, and N. Xiong, "Steganalysis of LSB matching using differences between nonadjacent pixels," *Multimedia Tools and Applications*, vol. 75, no. 4, pp. 1947–1962, 2016.

[19] Z. Xia, X. Wang, X. Sun, and B. Wang, "Steganalysis of least significant bit matching using multi-order differences," *Security and Communication Networks,* vol. 7, no. 8, pp. 1283–1291, 2014.

[20] Z. Xia, X. Wang, L. Zhang, Z. Qin, X. Sun, and K. Ren, "A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 11, pp. 2594–2608, 2016.

[21] C. Yuan, X. Sun, and R. Lv, "Fingerprint liveness detection based on multi-scale LPQ and PCA," *China Communications*, vol. 13, no. 7, pp. 60–65, 2016.

[22] C. Yuan, Z. Xia, and X. Sun, "Coverless image steganography based on SIFT and BOF," *Journal of Internet Technology*, vol. 18, no. 2, pp. 209–216, 2017.

[23] J. Zhang, J. Shen, L. Wang, and H. Lin, "Coverless text information hiding method based on the word rank map," in *The Second International Conference on Cloud Computing and Security (ICCCS'16)*, pp. 145–155, Nanjing, China, 2016.

[24] Z. Zhou, H. Sun, R. Harit, X. Chen, and X. Sun, "Coverless image steganography without embedding," in *The First International Conference on Cloud Computing and Security (ICCCS'15)*, pp. 123–132, Nanjing, China, 2015.

[25] Z. Zhou, Y. Wang, Q.M. Jonathan Wu, C. Yang, and X. Sun, "Effective and efficient global context verification for image copy detection," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 48–63, 2017.

Biography

Jianjun Zhang works as an associate professor in Hunan Normal University, China, and is currently working towards the PhD degree in computer science and technology at the College of Computer Science and Electronic Engineering, in Hunan University, China. His research interests include network and information security.

Huajun Huang is currently a faculty member in the college of Computer and Information Engineering at Central South University of Forestry & Technology. His overall research area include of Webpage information hiding and hidden information detection, XML Watermarking, Anti-phishing, Mobile Device Forensics. Dr. Huang received his Ph.D. from Hunan University in 2007, M.S. degrees from Hunan University in Software Engineering (2004), and a B.A. in Applied Physics from Yunnan University (2001).

Lucai Wang received the BE degree in Hunan University, China, in 1990, the PhD degree in Electronic Information Engineering from Hunan University, China, in 2006. He works as a professor in the College of Engineering and Design, Hunan Normal University. His research interests include intelligent information processing.

Haijun Lin received the BE degree in Hunan Normal University, China, in 2004, the PhD degree in Electronic Information Engineering from Hunan University, China, in 2009. He works as an associate professor in the College of Engineering and Design, Hunan Normal University. His research interests include intelligent information processing.

Deng Gao is currently a faculty member in the college of Software at Hunan Vocational College of Science and Technology. She received her M.S. degree from Central South University in Software Engineering in 2016. Her research interests include data mining and semantic networks.