Coverless Steganography Based on English Texts Using Binary Tags Protocol

Yulei Wu[1,2,3], Xianyi Chen[1,2,3], Xingming Sun[1,2,3]
[1] Jiangsu Engineering Center of Network Monitoring,
Nanjing University of Information Science and Technology, China
[2] Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology,
Nanjing University of Information Science and Technology, China
[3] School of Computer and Software, Nanjing University of Information Science & Technology, China
wyl940314@hotmail.com, {0204622, sunnudt}@163.com

Abstract

Information hiding is an evolving and challenging research field. Lately, a novel information hiding method that utilizes the original cover without modification to transmit secret messages has raised attention of the scholars. This method, coverless information hiding, can resist various kinds of existing steganalysis techniques due to its property. In terms of text steganography, coverless information hiding methods have made headway, most of which aim at Chinese steganography. Thus, this paper proposes a coverless steganography based on English, using the tag based protocol to make the secret communication possible. Experimental results reveal that the proposed method has decent success rate and hiding capacity.

Keywords: Coverless information hiding, Tag-based protocol, Binary, Steganography

1 Introduction

Information hiding is a research filed with a long history [1-5], and steganography, a technique that transmits the seemingly ordinary text or multimedia files with embedded information (also known as covers) to carry out imperceptible communication, is an important branch of it. In conventional information hiding methods, the senders usually conceal the secret information through modifying the attributes of natural covers or generating the artificial ones. However, with progress of technology, none of these methods can resist the steganalysis [6-8]. Coverless information hiding emerges as the times require.

The major difference between coverless information hiding and the conventional one is that coverless information hiding methods take advantages of the attributes of the natural covers rather than modify them or generate them. All the stego-covers used in the communication are actually unmodified natural files, making this technique able to resist various kinds of steganalysis. This characteristic makes coverless information hiding fleetly become a popular and challenging research field. Due to non-modification of the stego-cover, the receivers cannot extract the secret messages through conventional methods. Therefore, the information extraction phase of the coverless information hiding is different from the conventional ones. To extract the correct messages concealed in the received files, two sides of the communication ought to establish a communication protocol beforehand. The senders retrieve the normal files in big data according to the protocol, and the receivers understand the information correctly via it. According to different types of the stego cover, the coverless information hiding can be divided into various classes. Zhou put forward an image-based coverless information hiding scheme in [9], Yuan improved the scheme by using SIFT and BOF features [10]. Chen proposed a text-based steganography in [11].

The text-based coverless information hiding has made some achievements, but most of them are limited to the Chinese language. In terms of other languages, some researchers put forward a few preliminary ideas without specific solutions [12-13]. This paper makes an attempt in the coverless information hiding field of western languages, proposing a steganography based on English texts.

The remainder of the paper is organized as follows: 3 text-based coverless steganography will be introduced in Section 2, the proposed method will be presented in Section 3. The following part is the experimental results and analysis. Afterwards, the conclusion is put forward in the final section.

2 Related Work

In this section, the tag-based protocol will be briefly introduced first. And three Chinese coverless information

*Corresponding Author: Xingming Sun; E-mail: sunnudt@163.com

DOI: 10.3966/160792642018031902028

hiding methods will be reviewed.

2.1 Tag Based Communication Protocol

As is mentioned above, the senders only send the natural texts to the receivers, so the protocol is required in order to help the receivers extract the information. The core part of the protocol is tag, which is used to locate secret message or part of it (also called as keyword). The tag is usually converted from parts or combinations of the characters. To improve the security of the secret communication, the tags used in each communication must be different. The protocol is utilized to choose the tag sets which used by the communicating parties in advance. Both sides share the same tag sets in each communication to transmit the information. The senders retrieve files with certain tags so that the receivers can obtain the messages with them. Different tag selection strategies are applied to different protocols.

2.2 Steganography Based on Chinese Mathematical Expression

In [14], Zhou employed the Chinese mathematical expression as the location tag. His main idea is to use Chinese mathematical expression to convert each single character into parts, different parts are used as different tags. One Chinese character is divided into different tags, and the characters following the original character is the part of the secret message. By using this method, one or two characters can be concealed in each text. The receivers extract all words via the tags and combine them into the actual secret message.

2.3 Steganography Based on Chinese Character Encoding

In [15], Chen used an encoding technology of Chinese characters to convert the Chinese into bit stream, and utilized 6 successive bit as the tag. The major point is to translate the Chinese characters into UNICODE first, and then use the odevity of the UNICODE to decide the converted value of a character. If the UNICODE of a character is odd, this character will be converted into 1 and vice versa. To improve the hiding capacity, phrases instead of characters are utilized as keywords here.

2.4 Steganography Based on Hybrid Tags

We proposed a method using both tag sets mentioned above to improve the performance. The tag protocol used in this method is formed with two separate sub-protocols, one is Chinese mathematical and the other is Chinese encoding. In the information hiding phase, both tag conversion methods are involved. The senders segment the secret message into phrases and search for texts that satisfy the demands according to two sub-protocols. Then the optimal stego-cover sets will be transmitted to the receiver. This method improved both the success rate and the hiding capacity in a certain degree.

3 The Proposed Method

The features of coverless information hiding that none but the unmodified natural texts are sent to the receivers make the methodology of it completely different from that of the conventional ones. In coverless steganography, the information hiding is actually the process of retrieval. Therefore, the essential part of it is the database. After establishing the database, the pre-processing and index of the texts are also required to make the algorithm work more efficiently. Through searching for the texts that satisfy the requirements according to the shared protocol, the senders manage to transmit unperceived messages in secret to the receivers with natural texts. The proposed method is divided into three phases, which will be introduced as follows.

3.1 Preparation Work

The preparation stage is the fundamental part of this algorithm. Three aspects of work are involved in this phase: the construction of the database, the design of the protocol and the index of the whole database.

3.1.1 Construction of the Database

As is mentioned above, the stego-texts transmitted in this method are unmodified texts. Therefore, a huge database consist of natural texts ought to be constructed. We search for the English novels which have the potential to become stego-texts on the internet. Besides, we negotiate with the NIST to get the Reuters Corpus Volume 1 (RCV1) [16-17]. To increase the quantity of texts, all the texts are divided into smaller pieces. The texts are segmented by paragraph in order to keep them meaningful and natural. The composition of the whole database is shown in the Table 1.

Table 1. The component of the database

Type of the database	Total size/GB	Amount of the texts
English Novels	0.3	~3.3 million
RCV1	0.5	~4.7 million

3.1.2 Design of the Protocol

Tag, a hint to locate the secret messages, is the key point of the communication protocol. In fact, tag is a sort of features which one or more English characters carry. And the features must satisfy the principle of randomness and universality in order to make it possible for different cipher-texts to be transmitted via different stego-texts. Referring to the Chinese coverless steganography, this paper proposes a tag selection method based on binary stream. There are 26 characters in summary in English language, the coding of the character is used to turn every character into digit. Every English article is converted from characters into binary stream, and each bit, 1 or 0 in digit, can correspond to the original character. The procedure of the conversion can be described in 2 steps:

Step 1. Pre-process the English article. The goal of the conversion is to turn an article into a binary stream with the same length as the count of the characters in this article. Therefore, the first step is to filter the noise in the article. The noise refers to the punctuation, such as comma ‘,’ and space ‘ ’. These punctuations will have an influence on the total count of ‘0’ and ‘1’ after conversion, breaking the balance between them.

Step 2. Translate the characters into binary numbers. After step 1, the article is turned into a string consisting only of English characters. Then this string will be translated into UNICODE expressed in 16-bit binary. The last bit of the UNICODE decides the actual digit of the character after conversion. An example is shown in the Figure 1.

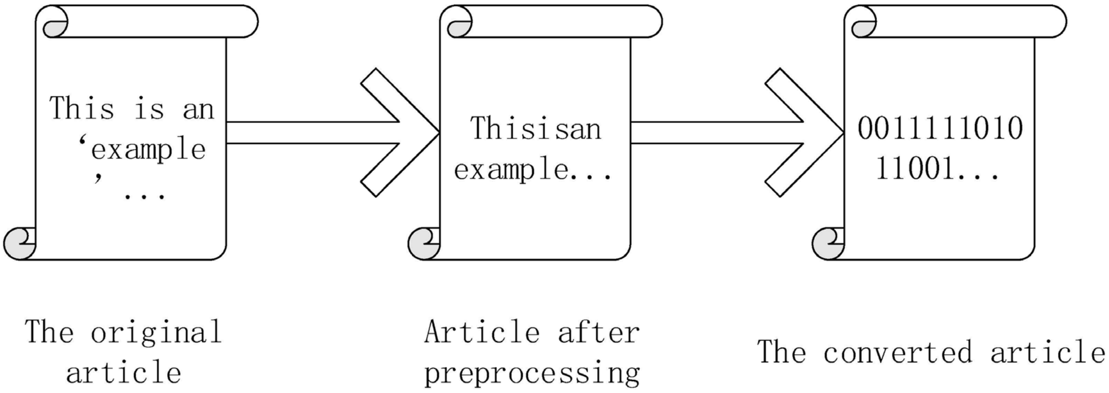


Figure 1. An instance of converting an article

After the conversion, 6-bit binary is selected as the tag. Therefore, the total count of the tags utilized is 64 (from 000000 to 111111). The tag sets used in the communication are determined by the protocol, and the simplified version of the protocol is presented as follows:

Firstly, before every communication, a secret key $K = I_1 I_2 \ldots I_n$ (n represents the total count of the key while I_i is a decimal number: 0-9) is generated by utilizing the identity of the communicating parties.

Secondly, the tag set $T = \{t_1, t_2, \ldots, t_n\}$ used in this communication is obtained via the secret key K and shared by the sender and the receiver.

In this paper, the K is a random number and the tag t_i is calculated as

$$t_i = t_0 + i * S \qquad \textbf{(1)}$$

(if $t_i > t_{\max}$, $t_i = \mathrm{mod}(t_i, N)$ $t_{\max}$ is the largest number of the tag, K is the total count of the tag, t_0 stands for the last but not digit of K while S is calculated as the last digit of K) for experiments.

Examples: Suppose K =500519, then t_0 =1, S =9, T ={1, 10, 19 … 54, 63, 9, 18}.

3.1.3 Index of the Database

The principle of the coverless steganography is retrieval rather than modification or generation. It is time-consuming to scan the texts on by on with such a large database, so the index of the whole database is necessary. There are couples of methods to retrieve articles over encrypted data [18-21]. When building the index, the tag as well as the word located by the tag (also expressed as keyword) ought to be recorded. In this paper, to improve the efficiency of the retrieving, the inverted index is employed. The process of indexing each article is described below:

Firstly, convert the article from English to binary using the rules mentioned above.

Secondly, scan the whole binary stream. Search for every tag (000000~111111) appearing in this stream and record their location. If the tag appears more than once, only the first appearance will be recorded.

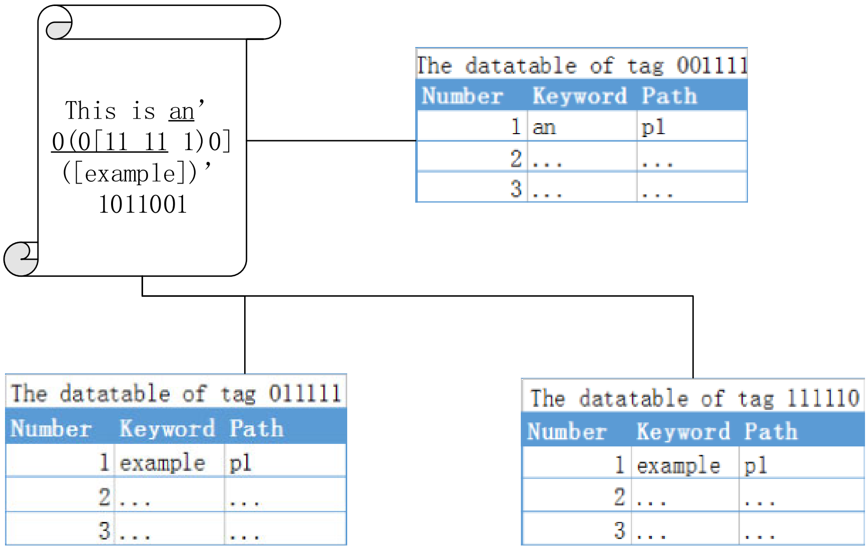
Thirdly, obtain the keyword string according to the location recorded. If the string located by the tag is not a complete word, take the first complete word after the string as the keyword. There is a one-to-one correspondence between one tag and one keyword.

Finally, draw data tables for each tag, in which the path of the article and the keyword that located by the tag is stored.

The example of the index and the data table is shown below in Figure 2.

3.2 Information Hiding

In the information hiding phase, the sender segments the secret messages into keywords at first. Then, he obtains the tag sets shared in this communication via the secret key according to the protocol. Afterwards, the combination of the tag and keyword will be utilized to retrieve the texts. The relationship between the keyword and tag is a one-to-one match. For each



The datatable of tag 001111		
Number	Keyword	Path
1	an	pl
2	...	...
3	...	...

The datatable of tag 011111		
Number	Keyword	Path
1	example	pl
2	...	...
3	...	...

The datatable of tag 111110		
Number	Keyword	Path
1	example	pl
2	...	...
3	...	...

Figure 2. An instance of how index works and the format of the data table

keyword, there will be a great quantity of texts that satisfy the requirements. However, only one text will be picked to be utilized as the stego-cover for this keywords. Finally, all the picked texts will be packed and sent to the receiver as the stego-cover. And the amount of cover-texts depends on the count of the keywords segmented from secret message. The pseudocode is shown in the Figure 3.

```
Input: Secret Message, Key
Output: Stego-cover texts
Begin:
(1) Segment the Secret Message $M$ by words, get the keyword set $\{m_1, m_2, m_3, ..., m_n\}$;
(2) Obtain the tag set $T = \{t_1, t_2, t_3, ..., t_n\}$ via the communication protocol;
(3) Combine each $t_i$ with $m_i$ to get the searching vector Set $\{c_1, c_2, ..., c_n\}\{c_i = t_i + m_i\}$;
(4) For each searching vector $c_i$:
(5) Retrieve the database for the articles $d_i$ that satisfy the requirements;
(6) Put all $d_i$ into candidate set $D_i$;
(7) For each $D_i$:
(8) Randomly pick an article $h_i$ as the stego-cover for this keyword;
(9) The combination of all selected texts form the stego-cover texts $H = \{h_1, h_2, h_3, ..., h_n\}$.
End
```

Figure 3. Pseudocode of information hiding

3.3 Information Extraction

Information extraction, the reverse process of the information hiding, allows receiver to obtain the original secret message with the key and stego-cover. The first step is to acquire the tag sets by employing the key along with the communication protocol. After that, all the stego-cover texts will be converted into binary stream with the same method using in the indexing phase. By searching for the first appearance of each tag in the corresponding article, the location of the keyword string will be achieved. Afterwards, the string located by this tag will be extracted with the following steps: (1) Extract the string with the help of the location. (2) If the string is not a complete word, the first word after the string will be taken as the keyword. One word will be extracted from one article after extraction procedure. Finally, combine all words to get the original secret message. The pseudocode is presented in Figure 4.

```
Input: Stego-cover texts, Key
Output: Secret Message
Begin:
(1) Obtain the tag set $T = \{t_1, t_2, t_3, ..., t_n\}$ via the communication protocol;
(2) Combine each $t_i$ with $h_i$ to get the extraction vector set $\{e_1, e_2, ..., e_n\}\{e_i = t_i + h_i\}$;
(4) For each vector $e_i$:
(5) Convert $h_i$ into binary stream;
(6) Search for the first appearance of the $t_i$ to get the location of the keyword;
(7) Obtain the keyword string $m_i$ located by $t_i$
(9) Combined all the extracted word $\{m_1, m_2, m_3, ..., m_n\}$ in sequence to obtain $M$.
(1) Obtain the tag set $T = \{t_1, t_2, t_3, ..., t_n\}$ via the communication protocol;
(2) Combine each $t_i$ with $h_i$ to get the extraction vector set $\{e_1, e_2, ..., e_n\}\{e_i = t_i + h_i\}$;
End
```

Figure 4. Pseudocode of the information extraction

4 Experimental Results

The performance of the proposed method is determined by a number of factors. Therefore, several experiments, including tests on the success rate as well as the hiding capacity, were carried out in this section to demonstrate the efficiency of the algorithm.

4.1 Success Rate

The information hiding strategy of the proposed method is totally different from that of the conventional ones, which causes that the embedding of the secret messages is not always successful. If the articles that meet the requirements of the protocol don't exist, the embedding of this secret message will fail. Therefore, the success rate of embedding has a great influence on the performance.

4.1.1 Tests on Different Quantities of Tags

The tag designed in the proposed method is a 6-bit binary (from 000000 to 111111). Therefore, the total count of the tags is 64. The amount of the recording entries in each tag table after indexing is shown in the Figure 5.

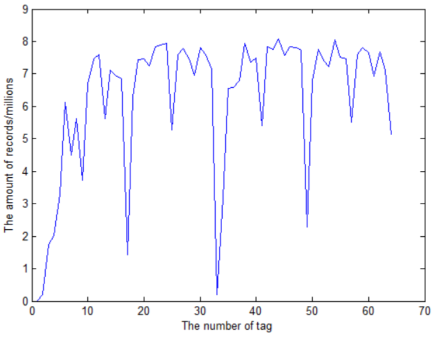


Figure 5. The amount of records of different tags

According to the figure, the line graph of amount fluctuates. The total amount of the tag with multiple 0 (e.g. "000000") is far smaller than others. The cause of this phenomena is possibly due to the restriction of the English word-formation and the difference in the frequency of English characters.

The tag is distributed unevenly in the articles, so the tag set utilized in the algorithm will influence the success rate of hiding. After indexing, the tags are sorted by the frequency of occurrence. In this experiment, the influence of different quantities of tags is tested. The success rate of top10, top20, top40, top 50 as well as all 64 tags are compared in the Table 2. The secret message set for the test is ENG which consists of around 500 25-word articles, ENG50 which consists of around 250 50-word articles and ENG100 which consists of around 100 100-word articles. The cover-text database is English Novels.

Table 2. Success rate of different tag set

Tag Set		Top 10	Top 20	Top 40	Top 50	All
Secret Message Set	*ENG*	97.7%	97.92%	97.92%	98.33%	89.28%
	ENG50	96.15%	97.01%	97.01%	97.01%	87.58%
	ENG100	91.24%	93.43%	94.89%	91.97%	80.34%

The success rate Σ is defined as:

$$\Sigma = \frac{Q}{J} \times 100\% \tag{2}$$

Where Q means the amount of the successfully hidden messages, and J denotes the total amount of the messages in the test set.

It is clearly shown in the chart that using all 64 tags will decrease the success rate by about 10 percent compared with other tag sets. The difference between the frequency of top 50 tags is subtle, making the success rate of the first 4 sets nearly the same. To strengthen the security of communication, the 4th set (Top 50) is chosen as the tag sets in the protocol.

4.1.2 Tests on Different Quantities of Data

The result of the first experiment also shows that the success rate of the algorithm is decent. However, compared to the Chinese coverless steganography with over 99% success rate, the rate is relatively low. What's more, a problem still exists-the rate of hiding a long message (ENG100) is not ideal. The lack of articles in database may be the root of this problem.

To improve the success rate and solve the problem above, a larger database-RCV1 is employed. The success rate of different database is tested with the same secret message sets used in the first experiment. The database is English Novels (EN for short) and the database which consists of both English Novels and 3 million articles of RCV1 (EN-RCV1_1 for short), as well as the English Novels added with the complete RCV1 (EN-RCV1_2 for short). The result of the experiment is recorded in the Table 3.

Table 3. Success rate of different database

Tag Set		EN	EN-RCV1_1	EN-RCV1_2
Secret Message Set	*ENG*	98.33%	98.54%	98.75%
	ENG50	97.01%	98.29%	97.29%
	ENG100	91.97%	93.05%	95.62%

It is indicated in the table that expansion of the database has a positive effect on the success rate. The more articles in the database, the higher the success rate is. What's more, the success rate of hiding long messages increases most significantly.

4.2 Hiding Capacity

Another important indicator of the performance is the hiding capacity (also known as embedding rate). Hiding capacity is the ratio of cipher-text length to the cover-text length. In this paper, the cove-text is a couple of meaningful articles, so the word count is utilized to calculate the capacity. The computational equation of capacity is defined as:

$$\Phi = \frac{\eta}{\psi} \times 100\% \qquad \textbf{(3)}$$

Where η stands for the total size of the secret message, and ψ represents the total size of all articles.50 messages from set ENG, ENG50 and ENG100 are used to calculate the hiding capacity. And the result is plotted in the Figure 6.

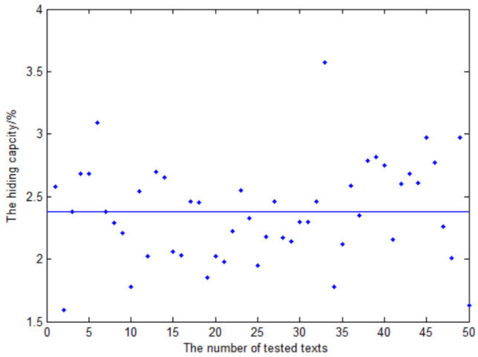


Figure 6. The figure of hiding capacity

4.2.1 Comparison with Conventional Steganography

The average value of the capacity is 2.38% approximately, which is plotted as a horizontal line in the figure. In Table 4 this method is compared with the similar text stegonography proposed in [18-23], from which we can see there is still a gap between this method and the conventional ones in terms of the capacity. However, the ability of anti-detection gives this method a unique advantage.

Table 4. Comparison between the proposed method and conventional ones

Type of methods	Average Capacity/%
Mimic Functions [1]	1.27
Component-based Chinese Watermarking [22]	2.17
Translation-based Steganography [23]	0.33
L-R and U-D representation [24]	3.53
Lexical [25]	0.5
Confusing [26]	0.35
Compression-based Steganogrpahy [27]	6.92
Proposed coverless Stegonography	2.38

4.2.2 Comparison with Coverless Steganography

The proposed method is also compared with the existing coverless information hiding method. Due to the difference in the calculation method of capacity, the data is transformed into the same format to make comparison. In [24], the proposed method utilized unmodified images as cover. The average size of the images in the database is about 100kb, and the average capacity is 1.86 Chinese characters per image. 1.86 Chinese characters occupy about 3.72 bytes, so after the transformation the hiding capacity is about 3.72÷100*1024≈0.36%. Similarly, the average size of the cover texts used in the coverless steganography based on Chinese is around 2kb, and the capacity is transformed in the same way. Detailed data is shown in the Table 5 below, from which it can be seen that the proposed method has the significant advantages in terms of capacity.

Table 5. Comparison between the proposed method and existing coverless ones

Type of methods	Average Capacity/%
Single-keyword method [11]	0.5
Multi-keyword method [14]	0.79
Method based on Chinese encoding [15]	1
Method based on Hybrid tags	1.04
Method based on Images [28]	0.01
Proposed coverless Stegonography	2.38

4.3 Security Analysis

The frequently-used information hiding techniques always transmit the modified or generated texts embedded with secret messages, making the cover-texts unable to resist the detection. As long as there are subtle changes or unnatural words in the texts, they can be easily detected with the developing steganalysis techniques. However, the cover-texts in the proposed method are all natural articles written by humans, which means almost all the existing steganalysis against the conventional information hiding methods cannot detect them. Suppose the sender wants to communicate with the receiver, all he need is to transmit some normal articles to the receiver. It is impossible for the third party to notice this because the normal communication like this is numerous. Even if the communication is noticed, the attacker has no access to the tag sets due to the randomness of the tag utilization, making the hidden messages avoid being extracted. To summarize, the features of the coverless information hiding make this technique hard to trace and detect, guaranteeing the security of the communication.

5 Conclusion

In this paper, a coverless text-based steganography using the natural texts to transmit secret messages is proposed. Compared with the commonly-used information hiding techniques, coverless information hiding has the ability to resist various steganography due to its features. The main idea of coverless information hiding is to search for unmodified texts according to the communication protocol. The receiver obtains the cover-texts with certain features. By extracting the features, the secret messages can be reformed. The text-based coverless information hiding has gained some achievements, but the proposed method is the first attempt in the English research field. The experimental results demonstrate the efficiency of the proposed method, but there's still plenty of space for improvement.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) (U1536206, 61502242, 61672294, 61602253, 61373133, 61232016, U1405254), Jiangsu Basic Research Programs-Natural Science Foundation (BK20150925), Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology (2014r026#), Open Fund of Demonstration Base of Internet Application Innovative Open Platform of Department of Education (KJRP1402), Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) Fund, Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET) Fund, National Ministry of Science and Technology Special Project Research GYHY201301030, 2013DFG12860, BC2013012.

References

[1] P. Wayner, Mimic Functions, *Cryptologia*, Vol. 16, No. 3, pp. 193-214, July, 1992.

[2] J. Brassil, S. Low, N. Maxemchuk, Copyright Protection for the Electronic Distribution of Text Documents, *Proceedings of the IEEE*, Vol. 87, No. 7, pp. 1181-1196, July, 1999.

[3] M. Mali, N. Patil, J. Patil, Implementation of Text Watermarking Technique Using Natural Language Watermarks, *IEEE International Conference on Communication Systems and Network Technologies*, Gwalior, India, 2013, pp. 482-486.

[4] W. Hsieh, N. Wu, C. Wang, A Novel Data Hiding Scheme for Binary Images with Low Distortion, *Journal of Internet Technology*, Vol. 11, No. 7, pp. 1057-1069, December, 2010.

[5] Z. Wang, Q. Mao, C. Chang, Q. Wu, J. Li, A Data Lossless Message Hiding Scheme without Extra Information, *Journal of Internet Technology*, Vol. 15, No. 4, pp. 657-669, July, 2014.

[6] Z. Xia, X. Wang, X. Sun, Q. Liu, N. Xiong, Steganalysis of LSB Matching Using Differences between Nonadjacent Pixels, *Multimedia Tools and Applications*, Vol. 75, No. 4, pp. 1947-1962, February, 2016.

[7] Z. Xia, X. Wang, X. Sun, B. Wang, Steganalysis of Least Significant Bit Matching Using Multi-order Differences, *Security & Communication Networks*, Vol. 7, No. 8, pp. 1283-1291, August, 2014.

[8] Z. Zhou, Y. Wang, J. Wu, C. Yang, 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, January, 2017.

[9] Z. Zhou, H. Sun, R. Harit, X. Chen, X. Sun, Coverless Image Steganography Without Embedding, *Lecture Notes in Computer Science*, Vol. 9483, pp. 123-132, January, 2016.

[10] C. Yuan, Z. Xia, X. Sun, Coverless Image Steganography Based on SIFT and BOF, *Journal of Internet Technology*, Vol. 18, No. 2, pp. 435-442, March, 2017.

[11] H. Sun, R. Grishman, Y. Wang, Active Learning Based Named Entity Recognition and Its Application in Natural Language Coverless Information Hiding, *Journal of Internet Technology*, Vol. 18, No. 2, pp. 443-451, March, 2017.

[12] J. Zhang, L. Wang, H. Lin, Coverless Text Information Hiding Method Based on the Rank Map, *Journal of Internet Technology*, Vol. 18, No. 2, pp. 427-434, March, 2017.

[13] X. Chen, H. Sun, Y. Tobe, Z. Zhou, X. Sun, Coverless Information Hiding Method Based on the Chinese Mathematical Expression, *Lecture Notes in Computer Science*, Vol. 9483, pp. 133-134, January, 2016.

[14] Z. Zhou, Y. Mu, C. Yang, N. Zhao, Coverless Multi-

keywords Information Hiding Method Based on Text, *International Journal of Security and Its Applications*, Vol. 10, No. 9, pp. 309-320, September, 2016

[15] X. Chen, S. Chen, Y. Wu, Coverless Information Hiding Method Based on the Chinese Character Encoding, *Journal of Internet Technology*, Vol. 18, No. 2, pp. 313-320, March, 2017.

[16] D. Lewis, Y. Yang, T. Russel-Rose, F. Li, RCV1: A New Benchmark Collection for Text Categorization Research, *Journal of Machine Learning Research*, Vol. 5, No. 2, pp. 361-397, April, 2004.

[17] Reuters Corpora, http://trec.nist.gov/data/reuters/reuters.html, last accessed 2017/3/27.

[18] Z. Fu, K. Ren, J. Shu, X. Sun, F. Huang, Enabling Personalized Search over Encrypted Outsourced Data with Efficiency Improvement, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 27, No. 9, pp. 2546-2559, January, 2016.

[19] Z. Fu, F. Huang, X. Sun, A. Vasilakos, C. Yang, Enabling Semantic Search based on Conceptual Graphs over Encrypted Outsourced Data, *IEEE Transactions on Services Computing*, Vol. PP, No. 99, pp. 1-11, October, 2016.

[20] Z. Xia, X. Wang, X. Sun, Q. Wang, A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 27, No. 2, pp. 340-352, February, 2016.

[21] Z. Xia, X. Wang, L. Zhang, Z. Qin, X. Sun, 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, November, 2016.

[22] X. Sun, G. Luo, H. Huang, Component-based Digital Watermarking of Chinese Texts, *Proc. 3rd International Conference on Information Security*, Shanghai, China, pp. 76-81, 2004.

[23] R. Stutsman, C. Grothoff, M. Atallah, K. Grothoff, Lost in Just the Translation, *Proc. 2006 ACM Symposium on Applied Computing*, Dijon, France, pp. 338-345, 2006.

[24] Z. Wang, C. Chang, C. Lin, M. Li, A Reversible Information Hiding Scheme Using Left-right and Up-down Chinese Character Representation, *Journal of Systems & Software*, Vol. 82, No. 8, pp. 1362-1369, August, 2009.

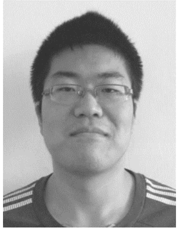
[25] Lexical Steganography, http://web.mit.edu/keithw/tlex/.

[26] M. Topkara, U. Topkara, M. Atallah, Information Hiding through Errors: A Confusing Approach, *Proc. SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents*, San Jose, CA, 2007.

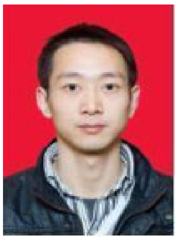
[27] E. Satir, H. Isik, A Compression-based Text Steganography Method, *Journal of Systems and Software*, Vol. 85, No. 10, pp. 2385-2394, October, 2012.

[28] Z. Zhou, Y. Cao, X. Sun, Coverless Information Hiding Based on Bag-of-Words Model of Image, *Journal of Applied Sciences*, Vol. 34, No. 5, pp. 527-536, September, 2016.

Biographies



Yulei Wu received the BE degree in software engineering from the Nanjing University of Information Science & Technology in 2016, China. He is currently working towards the MS degree in computer science and technology at the College of Computer and Software, in Nanjing University of Information Science & Technology, China. His research interest includes network and information security.



Xianyi Chen received his Ph.D. in Computer Science and Technology from Hunan University, China, in 2014. He is a lecturer in the School of Computer and Software, Nanjing University of Information Science& Technology, China. His research interests include information hiding, digital forensics and cloudcomputing security.

Xingming Sun is a professor in the School of Computer and Software, Nanjing University of Information Science and Technology, China. He received the B.S. degree in Mathematical Science from Hunan Normal University and M.S. degree in Mathematical Science from Dalian University of Technology in 1984 and 1988, respectively. Then, he received the Ph.D. degree in Computer Engineering from Fudan University in 2001. His research interests include information security, network security, cryptography and ubiquitous computing security