Label Model Based Coverless Information Hiding Method

Zhangjie Fu[1,2], Hongyong Ji[1], Yong Ding[2]

[1] Department of Computer and Software, Nanjing University of Information Science and Technology, China
[2] Guangxi Key Laboratory of Cryptography and Information Security, Guilin University of Electronic Technology, China
wwwfzj@126.com, z2273812251@gmail.com, stone_dingy@126.com

Abstract

Information hiding technology guarantees the security of information exchange in the internet. A highly secure information hiding method was presented recently -- Coverless information hiding and it attached many attention. But the low hiding capacity limits its application. A modified coverless information method called label model based coverless information hiding method was proposed in this paper. The new label model of the method use a tag to positioning the keywords as much as possible, and add a header file to hide the number of the keywords. The new method will hide more information compare with the coverless information hiding method before, because we usually hide fixed number of the information in the past. The experimental results show that the proposed method can improve the capacity of the method obviously.

Keywords: Coverless information hiding, Big data, Chinese mathematical expression, Word segmentation

1 Introduction

With more and more frequent information exchange in human society. Information security has become the common focus of both academy and enterprise. Based on this, large number of scheme was presented to prevent people from leakage, tampering etc. One of those schemes is steganography, an important research topic in the field of information security. In the meantime, the technology has promoted greatly the development of steganalysis. The goal of steganalysis is to identify suspected data, to determine whether they contains hidden information. Steganography and steganalysis are complementary and influence each other [1-4].

A new information hiding technology has been presented--coverless information hiding. Compared with other method, coverless information hiding biggest advantage is it has a high degree of invisibility.

Coverless information hiding is a method, which select the proper natural text from the big data as information carriers. Then send the carriers to the receiver directly and the receiver extract the hide information from carriers. The coverless information hiding technology is different from the conventional information hiding technology, it utilize the natural text from the internet or other data set directly. Due to the natural text, the coverless information hiding technology can resist all the steganography detection. It is easily justified that steganalysis is based on the modification of the carriers, but the coverless information hiding method is adopted in the natural text without modified.

Formidable imperceptibility of the method benefit from its features: no embedding, no modification and anti-detection [6]. We pick suitable natural text to meet receiver's requirement and sent it to them. We don't change the content of the natural text in the serial. These features make the carriers is made up of words used in the daily life. And that is why coverless information hiding can outwit different kinds of steganalysis.

What text is appropriate and that decide how to extract the secret information by the receiver from the natural text. Next, We will explained the selection of the text: firstly we divide the secret information up into the keywords, then create the "location tag" which represent the characteristics of the receiver, looking for the text that meet the characteristic of the location tag + keyword as the plaintext finally. So the receiver can get the tag with the characteristics and the keyword could be position easily.

Many coverless information hiding method was proposed in recent years. A lot of text-based coverless information hiding method and image-based coverless information hiding method were proposed [5-12]. The image-based information hiding method is same as the text-based method. For those text-based information hiding method inherent the characteristic of the coverless information hiding. As a result, they also can resist steganalysis of various types. We will described some typical coverless information hiding techniques recently:

At first, the coverless information hiding technology is used in the application of the text information hiding.

*Corresponding Author: Zhangjie Fu; E-mail: wwwfzj@126.com
DOI: 10.3966/160792642018091905022

[6] Proposed the method that use the Mathematical Representation of Chinese Character as the tags to positioning the keyword. [7-8] add the capacity of the information hiding through the optimization of the tag + keyword, and we will continue to optimize it in this paper and let the tag positioning more keywords instead of the fixed number of keywords. In addition, [9] use the new customized character encoding by themselves as the tag, and rewarded with success in the capacity. [11] Proposed a new scheme used with the word rank map to replace the tag, but their scheme is imperfect in the extraction. Because of the synchronized function between the sender and the receiver,

In the aspect of the image hiding, [5] proposed the feasible scheme to realize the image coverless information hiding, [10] use the SIFT and the BOF character in the image to make the image coverless information hiding come true.

This paper presented a modified coverless information hiding method based on the innovative hiding model. Moreover, the experiment show new model hide more information in less carriers. Our contributions are as follows:

We design a new coverless information hiding method, the new scheme add the header file to let tag positioning keyword as much as possible. Therefore, the proposed method could hide more information.

In order to prove our method is feasible. We established a 28.9G index to experiment with it, and the result confirm the method does increase the hiding capacity.

The rest of this paper is organized as follows. We will introduced the related work in section 2. And expound the proposed scheme including the detailed of the header file in section 3. Experiment and analysis are given in Section 4. Section 5 concludes the paper and the future work.

2 Related Works

This section will introduced the related technology used in this paper:

2.1 Chinese Character Mathematics Expression

The Chinese character mathematics expression was proposed by Sun et al.. in 2002 [12-13]. The function of the Chinese character mathematics expression is to use a mathematical expression to express the Chinese characters. The goal is to simplify the application of the Chinese character. Sun et al. stipulated the six expressive construction of the Chinese character. We will give an outline of that below.

The Chinese character mathematics expression split the Chinese into one or a few component. Then build a unique map between the component and the number. Just like the relationship of the English word and the letters. Except for the difference that the structure of Chinese is not from left to right. The basic structured way of the Chinese is left-right, up-down, left-down, left-upper, right-upper, and whole enclosed respectively [6].

We can express the Chinese character easily with the Chinese mathematical expression, in this paper we select the appropriate components as the tag to positioning the keyword in the natural text. It is better than choose the word as the tag because the basic component is more random and widespread.

2.2 Inverted Index

The inverted index is a data structure storing a list of mapping from the data to the location in the database. A record of the inverted index include a property value and the address of the value. The inverted index can allow us to search the appropriate text from the database conventional and quickly.

The general establishment method of the inverted index is described below:

(1) Split the document into the word term

(2) Pretreat the word term with the repetitive word

(3) Build the inverted list base the term.

Compared with the general method, we split the document according to the Chinese character mathematics expression in the paper, and the details will revealed in section 3.1

3 Proposed Method

This article has presented a coverless information hiding method based on the label model. It is an improved method based on single keyword [9]. The capacity of this text hiding method is risen greatly compared with the original method, here is how it works and concrete realization the whole process of the coverless information hiding:

3.1 Index

In the aspect of the index constructing and maintaining, the new approach is not so different from the general inverted index established method:

(1) For a document F, we could get the path and the name of F

(2) We split F with the third-party Chinese participle tool, with the help of the tool, we get a list of word term: t_1, t_2... t_m

(3) For the t_1, we add the special tag ‘start’, and the ‘start’ + t_1 is the first record of the document F

(4) For the t_i ($i > 1$), we split the t_{i-1} with the Chinese character mathematics expression and get the tags: tag_i, tag_{i+1}..., at the same time, the t_i is the candidate keyword then we will get some record like tag_i + $keyword_i$, tag_{i+1}+$keyword_i$...

Following the tag + keyword model design, we built a very large-scale index. The Figure 1 is structure

diagrams:

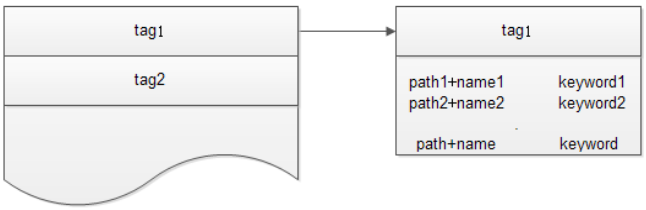

Figure 1. Structure of the index

We will take Figure 2 as a specific example to illustrate the process of the build index:

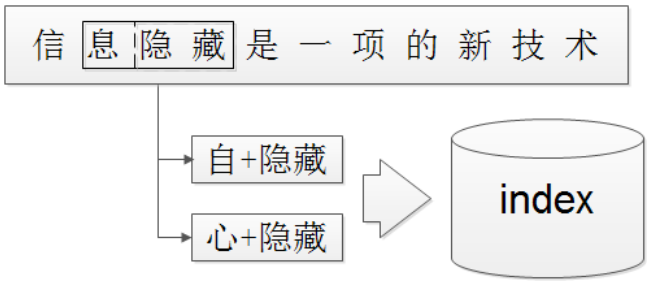

Figure 2. Example of the index build

If we prepared to build index for sentence T, we divide it into the keywords sequence: K_1, K_2...K_n. take K_1 as the first index entry, and attach a location tag L_1 before the keyword K_1, and here is the index structure: 'L_1 + T document name and path address + K_1'. For normal usage K_n, take the last word's Chinese character mathematics expression of K_{i-1} as the tag of K_i in the sentence T. the index is "L_{i-1}+ T document name and path address + K_i".

3.2 Hide

The improvement of capacity lies in the hiding method. In previous methods we hidden the secret information in the text, in order to extract information from the natural text possibly, we limit each tag positioning only one keyword. In the new way we allow the tag positioning a list of word as much as possible, as for the actual solving part of the extraction, we add a natural text to hide a header text. The header will tell receiver the number of the word in the next n file. This Figure 3 is the flow chart and details below:

(1) Pretreatment: for simplicity, we split the original text into terms, delete the stop word, and get the keyword sequence (k_1, k_2...k_n). We can add the keywords convert protocol to enhance the security of the system and to avoid directly appearance of the original text.

(2) Tags generation: the tag is the mapping to the receiver's character. This paper used the Chinese character mathematics expression as the tags. We generate the tags from the random tag list according to the parameters calculated by the receiver's character.

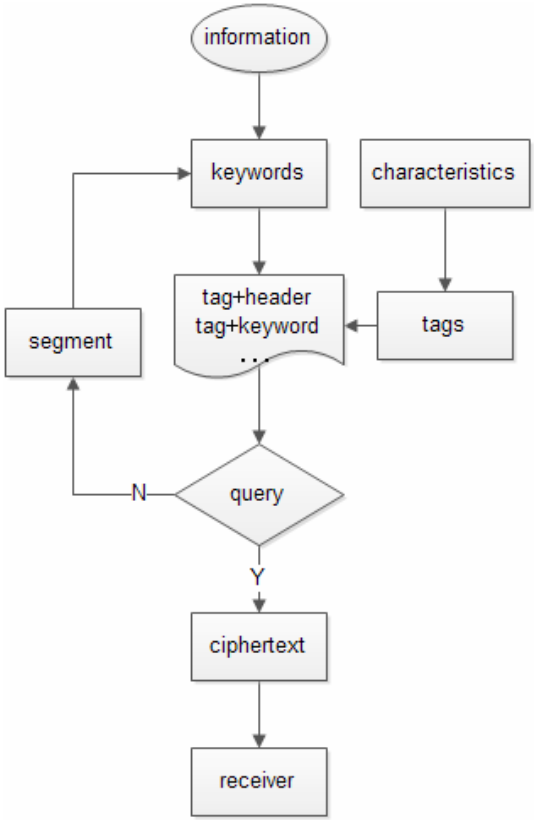

Figure 3. Information hiding

(3) information hiding: :we can acquired keyword sequence (k_1,k_2,...k_i) and tag sequence (t_1,t_2...t_n) from the two steps above. We use an example to illustrate the complete information hiding process.

For example, if we prepared to hide a sentence S, split S into a list of keyword K= {k_1, k_2...k_i}, and send the S to someone, we calculated the tags T= {t_1, t_2...t_i} by the character of someone with the customized function.

If we use the previous text coverless information hiding to hide K with T [6], we query appropriate text contain T+K = {t_1+k_1, t_2+k_2...t_i + k_i} from the index. As a result, we use i text to hide i word if all the requests are success

But in the proposed scheme, we provide a header file to help the next n file storage more keywords. Such as if we can find a natural text contain T+K = {t_1+k_1k_2, t_2+k_3, t_3+$k_4k_5k_6$...}, so we can query a natural text contain T+H (header file) = {T+ function (2, 1, 3 ...)} as the header file. The parameter of the function represent the number of the keyword in the next n file.

It can be seen that we use a header file and 3 natural text total 4 text to hide 6 keyword. It's clearly characterized by the advantage to use the proposed new scheme. More than that, we can just let the tag to position more and more keywords, it is possible with the expansion of text database.

3.3 Header and Parameter n

We will introduce the structure of the header and the setting of the parameter n in this part. We set the tags

to positioning the keyword as far as possible, but we can't always get what we want. We selected natural text compare to the tags from text library can only hide average two keywords generally. So we take two bit to hide the number of the additional keyword and the relationship is shown in Figure 4.

bit	num of word
00	0
01	1
10	2
11	3

Figure 4. Map of the relationship

The next step will show you how to decide the value of the parameter n. Previous research has shown that hiding success rate of individual tag positioning character goes up to 98%. The sum of the Chinese character is 91251, but the Chinese characters commonly used in about 2500 and Primary words in GB 2312-80(GB/T 16-55.1980) (total 3755)we added it to 4096 according to the frequency of the word in the text database. We can built a mapping between the Chinese character and the 12 bit (4096) binary. A natural text takes 2 bit of the header to tell the additional information to the receiver. So the value of the parameter n is 12/2=6.

All in all, we divided the carriers into one unit per 7 natural text (a header of additional information and six natural text of keyword).the structure of the carriers if shown in Figure 5:

Figure 5. Structure of the index entry

3.4 Extraction

The information extraction is the inverse procedure of information hiding. The tag generate parameters can be decided by the receiver's character, and the tag sequence could be calculated according to the tag generate parameters. Then positioning the location of keywords and the header according to the tags. At the last, we take the keyword by the header. The structure and the extraction algorithm is shown in Figure 6 below:

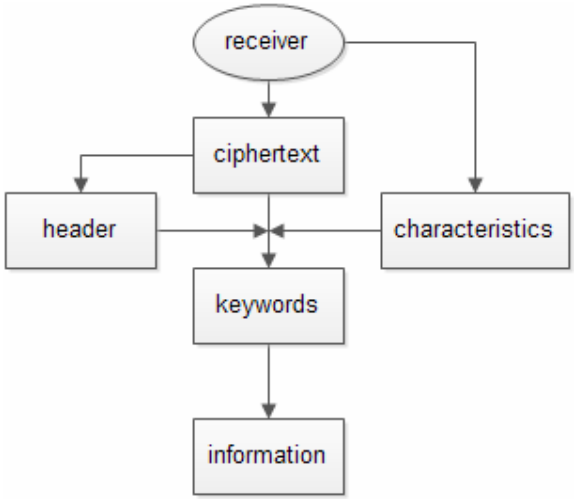

Figure 6. Information extraction

Information extraction ()

Input:

Cipher text sequence S(S_1,S_2 ...S_n) and Tag sequence T (T_1,T_2 ...T_m) the parameter n

Output:

Keyword K (K_1,K_2 ...K_s)

1. Split the S into Len(S)/(n+1) group
2. For each group:
3. For i from 1 to n+1:
4. If i is 1:(header file)
5. Query the position of the T_i, get next word W after the T_i, transfer W in binary. W' mean s the Len of the keyword in the next n file
6. Else :(keyword file)
7. Query the position T_i, get the next the word according to the W' as the keyword
8. End if
9. End for
10. End for

4 Experiment

We conduct the experiment and the method is proved to be correct and feasible. We use PyCharm community edition to investigate the performance of the proposed scheme. The index we build is appropriately a max of 28.9 Gigabyte. The plaintext we prepared for hiding is natural text crawled from the internet.

4.1 Success Rate

We test the scheme with 100 random natural news picked from the internet. Assume that the information is split into a list of words, the length of the word list is L, then we will hide tag + word, and the number of the failed hide is E. Therefore, the success rate R is defined as Eq. (1) and the result is shown below figure:

$$R = 1 - \frac{E}{L} \quad \textbf{(1)}$$

In the Figure 7, the columns represent the size of the news, and the size is ordered before. The scatter plots describe the hide success rate of each news. In the diagram, we use a line to marking the average success rate for analysis.

As can be seen clearly from the figure, the success rate fluctuate around the 95%. It's observed that our new method is highly capable for the brief news. In addition, we can find a news of a low hide success rate. We postulate that that news contain more rarely words, so this is reason for the low hide success rate.

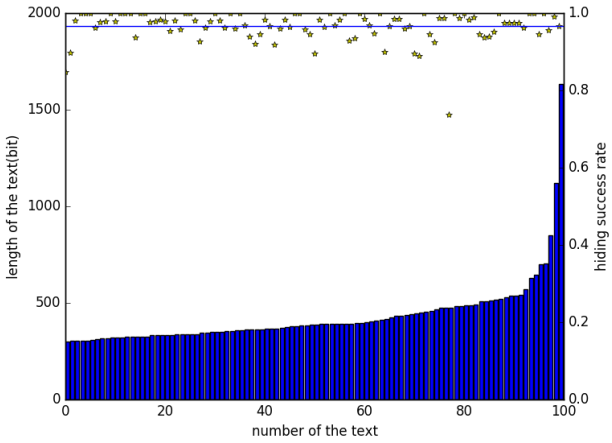

Figure 7. Success rate

Comparisons of the analysis and experimental data shows that the success rate of our new method still have some weaknesses. This is expected. Because in previous scheme. When we query the tag and the keywords from the database failed, we split the keywords to the single word, and query the tag + word one by one. This make the other scheme's success rate be the same as Chen's method.

But in this new method, instead of pursuing one hundred per cent of the success rate, we would like to hide more information in certain tag if possible. In the method, we won't segment the keyword for higher success rate. Because it will reduce the hiding capacity.

4.2 Data Hiding PerforMance

In this part, we will measure the information hiding capacity of our method. We assume that the size of the file prepared for hide is S. The number of the natural text contains header information we cost is H, the number of the query result meet the character of tag + keyword is T. Then the calculation of the hiding capacity C is shown as the below Eq. (2)

$$C = \frac{S}{H + T} \tag{2}$$

Not only that, we also extra 100 random file to hide, and compare to other coverless information hiding methods. The detail information of the hiding capacity is shown below:

In the figure 8, the green line represent the average hiding capacity of the [6] is 1 word per text, and the red line represent the average hiding capacity of the [8] is 1.57 word, then the yellow line represent the capacity of the [9] is 2.07word per text The blue stars represent the practical the hiding capacity in the experiment of the 100 text and the blue line is the average of our results: 2.5243 word per text. It is easy to see that the hidden capacity of our new method is better than the previous method.

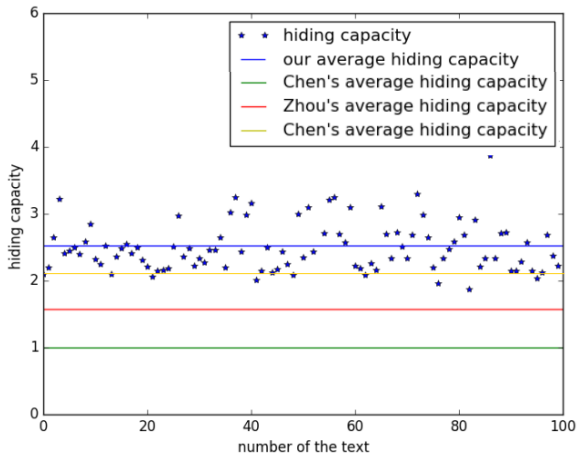

Figure 8. Hiding capacity

5 Conclusion and Future Work

This paper proposes a coverless information hiding method based on label model. We use a header file to hide additional information and six natural text to hide secret information as much as possible per text. This method inherited the feature of the coverless information hiding, so it has a strong resistance to all steganalysis method and human eyes. Finally, the experimental results show that this method can hide more information than the coverless information hiding method before.

In the future, we will try to use more tags to positioning more information in signal text instead of a signal combination of the tag + keyword. We also can add natural text purposeful for the missing combination of the tag + keyword. This will increase the success rate of the information hiding.

Acknowledgments

This work is supported by the NSFC (61772283, 61672294, U1536206, 61502242, U1405254, 61602253), BK20150925, R2017L05, PAPD fund, Guangxi Key Laboratory of Cryptography and Information Security (No. GCIS201713), Major Program of the National Social Science Fund of China (17ZDA092), Qing Lan Project, Meteorology Soft Sciences Project, National Science Foundation Grant (61772150), planning fund project of ministry of education (12YJAZH136) and China password

development fund (JJ20170217).

References

[1] J. Li, Y. Zhang, X. Chen, Y. Xiang, Secure Attribute-based Data Sharing for Resource-limited Users in Cloud Computing, *Computers & Security*, Vol. 72, pp. 1-12, January, 2018. doi: 10.1016/j.cose.2017.08.007

[2] Z. Huang, S. Liu, X. Mao, K. Chen, J. Li, Insight of the Protection for Data Security under Selective Opening Attacks, *Information Sciences*, Vol. 412-413, pp. 223-241, October, 2017.

[3] Y. Zhang, X. Chen, J. Li, D. S. Wong, H. Li, I. You, Ensuring Attribute Privacy Protection and Fast Decryption for Outsourced Data Security in Mobile Cloud Computing, *Information Sciences*, Vol. 379, pp. 42-61, February, 2017.

[4] Z. Fu, X. Sun, Q. Liu, L. Zhou, J. Shu, Achieving Efficient Cloud Search Services: Multi-keyword Ranked Search over Encrypted Cloud Data Supporting Parallel Computing, *IEICE Transactions on Communications*, Vol. E98.B, No. 1, pp. 190-200, January, 2015.

[5] Z. Zhou, H. Sun, R. Harit, X. Chen, X. Sun, Coverless Image Steganography without Embedding, *Proceedings of International Conference on Cloud Computing and Security*, Nanjing, China, 2015, pp. 123-132.

[6] X. Chen, H. Sun, Y. Tobe, Z. Zhou, X. Sun, Coverless Information Hiding Method Based on the Chinese Mathematical Expression, *International Conference on Cloud Computing & Security*, Nanjing, China, 2015, pp. 133-143.

[7] 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.

[8] 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.

[9] Z. Zhou, M. Yan, C.-N. 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.

[10] 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.

[11] 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.

[12] J. Zhang, J. Shen, L. Wang, H. Lin, Coverless Text Information Hiding Method Based on the Word Rank Map, *International Conference on Cloud Computing & Security*, Nanjing, China, 2016, pp. 145-155.

[13] X. Sun, H. Chen, L. Yang, Y. Y. Tang, Mathematical Representation of a Chinese Character and its Applications, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 16, No. 6, pp. 735-747, September, 2002.

Biographies

Zhangjie Fu received his Ph.D. in computer science from the College of Computer, Hunan University, China, in 2012. He is currently an associate professor at the College of Computer and Software, Nanjing University of Information Science and Technology, China. His research interests include Cloud & Outsourcing Security, Digital Forensics, Network and Information Security. His research has been supported by NSFC, PAPD, and GYHY. Zhangjie is a member of IEEE and a member of ACM.

Hongyong Ji received his B.S. in mathematics from Yancheng Teaching University, China, in 2014. He is currently pursuing his M.S. in computer science and technology at the Department of Computer and Software, Nanjing University of Information Science and Technology, China.

Yong Ding received the B.E. degree from sichuan university in 1998, the Ms degree and Ph.D. degree from Xidian university in 2003 and 2005 respectively. He is currently a professor with School of Computer Science and Information Security at Guilin University of Electronic Technology. He is interested in cryptography and information security from April, 2008 to September, 2009. His research interests include cryptography and information security.