Computer and Information Science; Vol. 10, No. 4; 2017
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

A Linguistic Steganalysis Approach Base on Source Features of Text and Immune Mechanism

Licai Zhu[1]

[1] Yancheng Teachers University, Yancheng, Jiangsu, China

Correspondence: Licai Zhu, Yancheng Teachers University, Yancheng, 2240000, China. E-mai: hobbyc@163.com

Received: Sepetember 19, 2017 Accepted: October 31, 2017 Online Published: October 31, 2017

doi:10.5539/cis.v10n4p60 URL: http://doi.org/10.5539/cis.v10n4p60

Abstract

linguistic steganalysis is a technique that discovering potentially hidden information embedded through using linguistically in plain text using. Varieties of syntax and multi-meanings of semantics for linguistics augment the difficulty of linguistic steganalysis intensely, thereby it is a challenge area. In this paper, we propose a novel steganalysis method for linguistics based on immune. This method has two attributions: i). basis statistical features of text are employed for blind steganalysis ii). immune technique is chosen to build a two-level detection mechanism to detect two categories of stego text respectively, one of which is Success-Stego-text and another is False-Stego-text. Appropriate detections are generated and preferable features are signed. Experiments prove the approach has higher accuracy than current steganalysis algorithms. Especially when the segment size of text is greater than 3kB, the accuracies of detecting for natural text and stego text are both more than 95%.

Key words: steganalysis; linguistic steganalysis; blind steganalysis; steganography; immune mechanism

1 Introduction

Compared with the research of image, audio and video hidden detection, (Kraetzer et al., 2015; Choubassi & Moulin, 2015; Kocal et al., 2016; Pankajakshan, Vinod & Ho, 2016; Bohme, 2016) text based information hiding detection is still a new and challenging field. Because of the widespread application of text communication media such as email, blogs and TXT on the Internet, researchers have accelerated the research on text steganography. There are three main types of text hiding detection: (Petitcolas et al., 1999), that is, format based, invisible character based and implicit detection based on natural language. The format of the hidden was measured through the analysis of (Li et al., 2016; Xiang et al., 2015) format text layout changes and kerning change based on text format is DOC, PDF etc. Hidden text detection based on invisible characters is aimed at the tag matching and invisible character changes in web language, such as (Huang et al., 2015; Huang et al., 2015; Huang et al., 2015), XML, HTML and so on. Compared with the previous two hidden text detection methods, hidden language detection based on natural language aims to discover the secret information loaded by text using linguistic knowledge. Because of the syntax diversity and semantic ambiguity, (Bennett et al., 2014; Mark et al., 2001; Chen et al., 2016) is difficult to obtain the feature for detection, which makes it a recognized research problem in the field of text hiding detection.

At present, only a few literatures have proposed a hidden detection technique based on natural language. The basic method is to analyze the features of the text, grammar, or semantics of the hidden text, and design special features to detect it. For example, Chen (Chen et al., 2016; Chen et al., 2016; Cuneyt, 2014; Forrest, 1997) were designed for different characteristic values were detected, such as the use of mutual information as the definition of the N window statistics, using the statistics obtained a mutual information matrix, the matrix as the standard, more natural text and hidden text mutual information with (Chen et al., 2016); The word positions of words are defined, and the word entropy and the word entropy of the word sets in the text are calculated by the word positions, and the (Chen et al., 2016) is tested by comparing the word entropy and the variation degree of the two texts; Compare the degree of separation and variation of word bits in text to detect (Cuneyt, 2014). Cuneyt M.Taskiran based on probabilistic context free grammar constructs 3 element Markov model, by comparing the words in a sentence such as the number of features to detect (Forrest, 1997).

The common problem of the above detection methods is that the design features cannot fully represent the text features. Therefore, when text fragments are longer, the detection performance is acceptable because of the difference between the natural text and the hidden text. But when the text fragments are short, the feature difference

is not obvious, and the detection rate is not high.

From the above analysis, we need to: (1) from the text element consideration for the detection of the characteristic, the reason is that although the hidden text in grammar, semantics and natural text similarity, but through the analysis of the text element, we found that the scope of their characteristics is different in many meta features. This is because the generated hidden text, choose random generally more natural text, making it smaller element changes. This feature is still satisfied when text fragments are relatively small. (2) the detection mechanism is reasonable, the reason is that although most element of hidden text will meet this characteristic, but there are still some characters are not satisfied, also there are redundant features of the phenomenon, so we need to choose the appropriate feature detection.

In this paper, a natural language hidden detection algorithm based on text meta feature and immune mechanism is proposed. This paper analyzes the difference between meta characters of natural text and hidden text. Then, the immune mechanism is used to train the detector to optimize the feature. The immune mechanism (Jacob et al., 2013) has been effectively applied to image hiding detection [20], and has achieved very good detection results. This mechanism generates a set of detectors that contain several features that separate the natural text from the hidden text. Experimental results show that the proposed algorithm has higher detection rate for hidden texts generated by three typical hidden tools.

The structure of this paper is as follows: in the second section, we give the basic idea of this method, including the analysis of the differences and the method flow of the selected meta features in the two types of text. The third section describes how to use the immune mechanism to train the detector to optimize the detection features. The details of the hidden text feature range and the process of using them to detect are explained in detail. In the fourth section, the experiment is compared with the current detection algorithm, and the problems which should be paid attention to are analyzed. Finally, the fifth section gives a conclusion.

2. Basic Idea of this Method

2.1 Comparison of Textual Meta Features

According to the linguistic knowledge of English, statistics of English texts can yield many relatively invariant statistical feature values, called meta features, such as space rates. This is because with the increase of the statistics, random words increase, the probability of each word is selected equal, makes character value tends to be stable. Therefore, when comparing the two sets of text sets, we find that their meta values do not change much. But this phenomenon is obtained through a large number of statistics of the text, the general single natural text, the main concern is the argument, the details of the content, so when comparing the two natural text, we found that the element values change greatly. The hidden text, due to the loading of secret information, choice of randomness, the element features of most values are relatively stable, the hidden features of text value between small changes, and because of the hidden requirements, most of its feature values range within the range of natural text. Therefore, it enables the use of smaller variation range of characteristics of hidden text to detect the existence of hidden text information, if there are multiple unknown text features are smaller in a given range, the text may be hidden text. We select 57 meta features and divide them into 5 cases as features to be selected. Since there is no literature to analyze the difference between natural text and hidden text by meta feature, this paper will analyze the meta feature in turn, and then give the comparison of two text element feature ranges.

The natural text and hidden text in different features:

(1) the average length of words

According to the principles of Zipf and Heaps, the average length of words is within a certain range. Hidden text, when loading secret information, the choice of words is very random, so there are fewer high-frequency words in the text, and the average length of words is longer. To increase the concealment of text and reduce the average length of words, the hidden tool increases the number of short words when creating hidden text, but this change may change the following features 2 and 3.Therefore, it can be seen as a balanced feature.

(2) space rate

According to the Heap principle, the rate of text space is approximately inversely proportional to the length of the word, and the punctuation is not taken into account when calculating. If the space appears two times or more, count only once. Obviously, the feature detection is more effective in identifying hidden algorithms that change the number of spaces used to hide information.

(3) the first N valid words and invalid words

According to the TF-IDF theory, text high-frequency words include effective high-frequency words (AFW) and

invalid high-frequency words (NFW). AFW is a valid keyword for expressing the topic of an article, while NFW is something without meaning, such as "the", "of", etc.

(4) percentage of letters

In English articles, when the amount of text is small, the distribution of text letters changes greatly. As the amount of text increases, the randomness of the letters increases, and the percentage of letters is almost unchanged. Random choice makes the hidden text when the text size is small, the change is less than the percentage of letters of natural text.

(5) percentage of initial letters

The distribution of the initial letters of English text is related to the distribution of the initial letters of a dictionary. Since the distribution of the initial letters in the dictionary varies little, the initial distribution of the text varies little. The randomness of the hidden text words natural text distribution is more stable.

2.2 Method Flow

Through the above analysis, we use the meta feature to represent the text, and then establish a reasonable detection mechanism to select the appropriate detection features. Here, it's important to note that when the hidden text fragment is large, its meta features are generally smaller;When the hidden text is small pieces (less than 3K), although most of the hidden text element is still in a small range (called the hidden text successfully (Success-Stego-Text SST)), but there are fewer hidden text, due to the loading of secret information, random choice of words was lower than that of SST, so many yuan characteristics their value is not in the range of SST feature (but still in the range of natural text element), we will this kind of text is called hidden failure text (Failure-Stego-Text FST).For these two types of hidden text, you need to test them in turn. Therefore, we establish a two level detection mechanism to detect two types of hidden text in turn.

The algorithm flow is as follows: first, the meta feature is extracted to represent the two type of text; Secondly, the feature range of SST is extracted firstly, and the SST detector and SST detection feature are trained by immune mechanism. Then, for the FST detected by the SST detector, the range of feature is extracted and the FST detector and FST detection feature are trained; Finally, the two detectors are constructed into two level detection mechanisms for detecting two hidden texts.

Since the algorithm does not need to consider whether there is hidden information or what method to hide, it has certain blind detection. The concept of blind detection is proposed by [28-30], which is used to detect image hiding. It refers to the method of detecting whether there is hidden information in the text without knowing the hidden way. The effect of the decision depends on the hidden, and different detection methods should be changed in detail in different hidden persons.

3. Training Process

The training process is divided into two stages, followed by the training of SST and FST detectors and their corresponding eigenvalues. In the first level of training, the input samples are hidden text and natural text that contain SST and FST. The training process is as follows: first, a primary detector is generated, then the detector is trained by the immune mechanism and the SST feature range to obtain the SST detector and the SST eigenvalue; In the two stage, the input sample is the natural text at the higher detection stage and the FST which is mistaken for the natural text. The training process is as follows: firstly, a two stage detector is generated, then the detector is trained by the immune mechanism and the FST feature range, and the FST detector and the FST eigenvalue are obtained. When detecting, the two detectors are combined into a two level detection system.

3.1 Related Definitions and Theorems

Set T (S) and T (N) to represent hidden text and natural text collections to be detected, T '(S) and' T '(N), respectively, indicating that detection is considered as a collection of hidden text and natural text.

Defines 1 hidden text detection rates SR and a natural text detection rate of NR

$$SR = |T(S) \cap T'(S)| \ / \ |T(S)| \quad (1)$$

$$NR = |T(N) \cap T'(N)| \ / \ |T(N)| \quad (2)$$

By equation 1 and 2, when SR increases, the NR decreases, and vice versa.

Define 2 feature groups F={f1, f2,... F57}, fi stands for characteristics. It can be considered as a chromosome gene. It has two values, vaule1 and Value2, with initial values of 1. They represent the ability to detect SST and FST. The greater the value, the stronger the detection ability of the feature. The detector C consists of several features that can be considered as chromosomes containing multiple genes. It has two parameters, SR and NR, which reflect

the detection capability of detectors, collectively referred to as fitness.

Defines 3 text affinity(*Affinity Score*)

Set text "X" and "Y", and remember "X.fi" and "Y.fi". The affinity is AS

$$AS= \sum_{i=1}^{57} coe[i] \times \frac{1}{1+\left|X.fi - Y.fi\right|},$$

Among them, COE [i] is the parameter of each feature, which is used to increase the division.Set the text T and the text set T, and the maximum affinity and minimum affinity for the text in T and T are:

MaxAS=maximum AS(t, each text of T)

MinAS=minimum AS(t, each text of T)

3.2 Training Process

The training sample T consists of the hidden text set T (ST) and the natural text T (NT). Split T into N groups, with the same percentage of hidden text and natural text in each group. Let Tk denote a set of samples, K and N. These samples will be used sequentially to train the detector set.

Level 1 training phase: the detector trained at this level is used to detect SST text. First, how to obtain the feature range of SST is described. Due to overfitting and the fact that FST makes changes in features too large, only approximate range of features can be obtained. A hidden set of text S is selected randomly and its characteristic range is counted as the feature range of SST. The maximum and minimum values of each feature are denoted as MaxS[i] and MinS[i], respectively, i=1... 57.

A collection of M chromosomes is generated randomly as an initial immune detector. Set higher SR and NR values to make Sr and Nr. The calculation of chromosome NR, SR or Sr requirements.

When cloning chromosomes, the selection mechanism is used to accelerate the convergence time, that is, to increase the selection probability of chromosomes whose fitness exceeds the average fitness. In order to maintain chromosome diversity, when the chromosome concentration exceeds the average concentration, the probability of selection is reduced. The single nucleotide mutation was used for gene mutation after cloning, and the mutation rate was 0.4.

According to the fitness sequence, select the top N chromosomes into memory as Base chromosome, chromosome value1 gene according to the new adjustment value, the number of value1 gene in the chromosome in memory by decision.

When the memory of chromosome fitness to meet the SR more than Sr and NR is not less than NR, the end of the algorithm. The memory chromosome represents the mature SST detector to detect the SST text; the value1 value represents the eigenvalues of the SST to generate a new SST detector.

Two level training stage: the text set after the last level detection is T ', which consists of the majority of the natural text T (N) and the minority of the hidden failure text T (FST).At this level, the main test is mistaken for natural text of the FST. To get the feature range of FST, you need to extract part of FST from T'. Specifically, select a set of natural text T and calculate its MinAS and MaxAS for each text in the text 'T'. For a large number of T (N), the numerical distribution of MinAS satisfies the normal distribution, while T (FST) is only a part of the hidden text, whose distribution does not satisfy the normal distribution, and can be verified by theorem 1.Therefore, the difference between T (N) and T (FST) and T 'MinAS is greater than their MaxAS difference, so MinAS is used to distinguish between the two. The distributions of these two types of text sets have significant differences, indicating that there are partial FST, which are very different from natural texts, which provides an optional space for FST. The selection criteria are FST that can distinguish more than 98% natural texts.

The training steps at this stage are similar to those of the previous training. The difference is that the Memory Base stores the FST detector and adjusts the characteristic values such as value2.

4. Experiment and Analysis

This section uses this algorithm to detect hidden text generated by three typical hidden tools and compare them with several current detection algorithms. The 4.1 section briefly introduces each hidden tool; the 4.2 section selects the hidden text generated by natural text and hidden tools as samples and makes a comparative experiment. The 4.3 section puts forward the problems that the algorithm needs to pay attention.

4.1 Three Typical Hidden Tools

Nicetext is a tool for masquerading ciphertext as a natural text, [31, 32].The tool preserves some writing techniques

and a dictionary in advance to hide text. It also gets writing styles and vocabulary from other natural texts.

Texto is a uuencoded or PGP ascii-armoured ASCII code can be translated into English for information hiding tool [33].The tool is used for conversion of binary data, especially ciphertext data. It can generate documents similar to the natural text of English and can resist simple mail filtering.

Markov-Chain-Based is a typical tool for generating text (Chen et al., 2016; Chen et al., 2016; Cuneyt, 2014; Forrest, 1997).It starts with the Markov signal source and creates a conversion chain to generate dense text like natural text

4.2 Experiment

Selected 9800 novels from sixteenth Century to twentieth Century, extracted from the first 5000 high-frequency words used to calculate P (AFW) and n (AFW).Then 773 famous novels, including Shakespeare's plays and Aesop's fables, are selected as samples of natural texts. Use the hidden tools to generate 422 text as hidden text samples, 212 of them NICETEXT, 110 texto, and 100 markov-Chain-Based text.

30 text films were extracted from each of the two categories, (773+422)*30= 35850.When calculating the characteristic range of SST, one of the selected samples is randomly selected to calculate its characteristics. Randomly select one from natural text as a natural text set for calculating FST. Then merge the two text slices into 30 sets of mixed sample fragments, each containing 2/3 natural text and 1/3 hidden text.

We compare the current detection algorithms with text, slice size, 20kB, 10kB, 5kB, 4kB, 3kB, 2kB, 1kB, 0.5kB, 100B, and 8 cases. In the experiment, 20 of the sample fragments were selected for training, and the remaining 10 samples were tested.

Table 2. Comparison of the algorithm with the current algorithm

Text Size	Our Algorithm		Previous Algorithm
	SR	NR	
20kB	99.99%	99.99%	98.50%
10 kB	99.99%	99.99%	95.51%(Cuneyt, 2014)
5 kB	98.97%	99.99%	92.95%(Cuneyt, 2014; Forrest, 1997)
4 kB	98.52%	99.82%	91.03%(Forrest, 1997)
3 kB	95.83%	95.48%	86.11%(Forrest, 1997)
2 kB	93.50%	90.09%	79.27%(Forrest, 1997)
1 kB	88.42%	86.81%	71.15%(Forrest, 1997)
0.5 kB	86.82%	82.03%	-
100B	85.35% 90.07%	48.36% 43.32%	84.9%(Jacob et al., 2013) 38.6%(Jacob et al., 2013)

According to table 2, the proposed algorithm outperforms the current detection algorithm. When the text size is not less than 3k, SR and NR of the recognition rate of more than 95%, when the text only a word, its size is approximately 100B, and the (Jacob et al., 2013) detection of the same length. When the algorithm is approximated by SR and (Jacob et al., 2013), its NR is 10% higher than that of (Jacob et al., 2013); when SR reaches 90%, NR is 5% higher than (Jacob et al., 2013).

5. Conclusion

In this paper, a new natural language detection algorithm is proposed based on the analysis of text features. The algorithm uses multiple element feature better expresses the text characteristics, and then through the analysis of changes of the features, the hidden text is divided into SST and FST, respectively, after feature extraction are designed, the corresponding detector, so the detection algorithm in the text is relatively short time also have a higher rate. Experiments show that when the length of text film reaches 3K, both SR and NR are more than 95%, and the stability is stronger.

Because the features used in this algorithm are the basic statistical characteristics of the text, and not for the special characteristic of hiding algorithm design, it can detect different hidden text hiding algorithm, blind detection algorithm has a certain. But because there is no more to consider change characteristics of the hidden text in the syntax and semantic level, so the algorithm in the detection of sentence NR is not high. Therefore, changes in these levels can be taken into account to identify new features.

At present, although there are still many challenges and difficulties in the research of hidden detection based on natural language, more and more attention has been paid to it. The key factor in the detection technique is the amount of load hidden. The algorithm presented in this paper shows that it is feasible to detect hidden text by using text meta features.

Acknowledgement

This work is supported by Prospective Joint Research project of Jiangsu Province under Grant No. BY2016066-05, Teaching reform project of Jiangsu Provincial Department of Education under Grant No. 2015JSJG526.

References

Bennett, K. (2004). Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text[M]. Purdue University, CERIAS Tech, Report 2004-13.

Bohme, R. (2016). Weighted stego-image steganalysis for JPEG covers. Information Hiding - 10th International Workshop[A], LNCS, 5284, 178-194.

Chand, V., and Orgun, C. O. (2014). Exploiting linguistic features in lexical steganography: Design and proof–of-concept implementation[A]. Proceedings of the 39th Annual Hawaii International Conference on System Sciences.

Chen, Z. L., Huang, L. S., Yu, Z. S., Li, L. J., & Yang, W. (2016). A statistical algorithm for linguistic steganography detection based on distribution of words[A]. 3rd International Conference on Availability, Security, and Reliability.

Chen, Z. L., Huang, L. S., Yu, Z. S., Yang, W., Li, L. J., Zheng, X. L., & Zhao, X. X. (2016). Linguistic Steganography Detection Using Statistical Characteristics of Correlations between Words[A]. Information Hiding - 11th International Workshop.

Chen, Z. L., Huang, L. S., Yu, Z. S., Zhao, X. X., & Zheng, X. L. (2016). Effective Linguistic Steganography Detection[A]. Proceedings - 8th IEEE International Conference on Computer and Information Technology Workshops.

Choubassi, M. E., & Moulin, P. (2015). Noniterative Algorithms for Sensitivity Analysis Attacks[J]. IEEE Trans. *Information Forensics and Security, 2*(3), 113-126.

Cuneyt, M. T. (2014). Umut Topkara, Mercan Topkara, Edward J. Delp. Attacks on lexical natural language steganography systems[A]. In Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents.

Forrest, C. B. (1997). Computer Immunology[M]. *Communications of the ACM, 40*(10), 88–96.

Huang, H. J., Sun, X. M., & Sun, G. (2015). Detection of Steganographic Information in Tags of Webpage[A], In Proceedings of The Second International Conference on Scalable Information Systems, ACM Press.

Huang, H. J., Sun, X. M., & Tan, J. S. (2015). Detection of Hidden Information in Tags of Webpage Based on Tag-mismatch[A]. Proc. & Conf. on Intelligent Information Hiding and Multimedia Signal Processing, IEEE Press.

Huang, J. W., Sun, X. M., Huang, H. J., & Luo, G. (2015). Detection of Hidden Information in Webpages Based on Randomness[A]. The Third International Symposium on Information Assurance and Security.

Jacob, T. J., Gregg, H. G., Roger, L. C., Jr, G. B., & Lamont, B. (2013). Steganography Detection Using a Computational Immune System: A Work in Progress[J]. *International Journal of Digital Evidence, 4*(1).

Kocal, O. H., Yuruklu, Emrah, & Avcibas, I. (2016). Chaotic-type features for speech steganalysis[J]. IEEE Trans. *Information Forensics and Security, December, 3*(4), 651-661.

Kraetzer, Christian, Dittmann, & Jana (2015). Pros and cons of mel-cepstrum based audio steganalysis using SVM classification[A]. Information Hiding - 9th International Workshop, LNCS, 4567, 359-377.

Li, L. J., Huang, L. S., Zhao, X.X., Yang, W., & Chen, Z. L. (2016). A statistical attack on a kind of word-shift text-steganography[A]. Proceedings-4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing.

Mark, C., George, I. D., & Marc, R., (2001). A Practical and Effective Approach to Large·Scale Automated Linguistic Steganography[A]. In proceedings of the 4th International Conference on Information and Communications Security, LNCS, 2200, 335-345.

Pankajakshan, Vinod, & Ho, A. T. S. (2016). Improving video steganalysis using temporal correlation[J].

Proceedings - 3rd International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 1, 287-290.

Petitcolas, F. A. P., Anderson, R. J., Kuhn, & Markus, G. (1999). Information hiding - a survey[J]. *Proceedings of the IEEE, 87*(7), 1062-1078.

Xiang, L. Y., Sun, X. M., Luo, G., & Gan, C. (2015). Research on Steganalysis for Text Steganography Based on Font Format[A], The Third International Symposium on Information Assurance and Security.