Advanced Engineering Forum *Online: 2012-09-26*
ISSN: 2234-991X, Vols. 6-7, pp 361-366
doi:10.4028/www.scientific.net/AEF.6-7.361

A Robust Webpage Information Hiding Method Based on the Slash of Tag

Yujun Yang[1,2,a], Yimei Yang[1,b]

[1]Department of Computer Science and Technology, Huaihua University, Huaihua, 418008, China

[2]School of Computer Science and Engineering, University of Electronic Science and Technology, Chengdu, 610054, China

[a]mlsoft163@163.com, [b]yym1630@163.com

Keywords: Slash; Information Hiding; Tag Attribute; Webpage.

Abstract. Nowadays,the information hiding technology is a hot spot in the field of information security, and is applied in many fields,such as digital multimedia copyright protection and secret communication. According to the analysis of the characteristics of browser in parsing HTML of the webpage and the little capacity available for information hided in webpage, a new robust webpage information hiding method with the slash of tag attributes has been proposed in this paper, which overcomes the shortcoming of the ability of imperceptibility and the ability of contradict with the machine filtration of traditional webpage information hiding algorithms and has greater embedded capacity than some other algorithm based on tag attributes. This method has good performances in invisibility and higher applied value as proved by the experiments.

Introduction

Information hiding[1] is to hide some secret information in innocuous-looking cover objects, such as audios, images, videos, texts, etc.. In recent years, Information hiding has generated significant research and commercial interest. The primary factors contributing to this surge are widespread use of the Internet with improved bandwidth and speed, regional copyright loopholes in terms of legislation; and seamless distribution of multimedia content due to peer-to-peer file-sharing applications.

HTML is a hypertext markup language for writing hypertext files, namely webpages, which are used to convey information through the Internet. With the development of the Internet as a main communicative means, webpages have enjoyed an extensive application in the Internet. Meanwhile, a wide variety of steganographic methods[2-4] for webpages have emerged. According to the analysis of the characteristics of browser in parsing HTML of the webpage, the source codes of a webpage are a plain text that contains small markup tags, by which the web browser is instructed how to display the page. Information hiding based on webpage uses a webpage as a cover, and then embeds some secret information into the source codes of the webpage, while the displaying effect will remain unchanged. Through analyzing the criterion of HTML and References [2-5], we have defined three main information hiding methods: 1) Based on the invisible characters embedding; 2) Based on the changing of letter upper and lower cases in tags; 3) Based on the changing the order of attribute tags pair.

The first two methods are information hiding methods based on document format. The method of embedding invisible characters is to embed extra invisible characters between tags, or after every row, or after the whole document, to encode secret information. The second method is based on the fact that letters in tags are always case-insensitive, therefore the cases of tag letters can be modified without changing the visible document or the file size. So, define the uppercase letter as the bit “0” and the lowercase letter as “1”, secret information can be embedded into a webpage by changing of the letters upper and lower cases in tags.

A new robust webpage information hiding method with the slash of tag attributes has been proposed in this paper, which overcomes the shortcoming of the ability of imperceptibility and the ability of contradict with the machine filtration of traditional webpage information hiding

algorithms and improves the embedded capacity of the other algorithm based on tag attributes. According to the embedded rule, firstly the sequenced tags entity set is acquired from the webpage. Then the message is encrypted by a two-value chaotic sequence generated by Logistic map system. The value format of a certain attribute in tags is selected and a modification is made to them based on the encrypted message, which is whether it has single quotation mark. The analysis shows that the method has good imperceptibility and perfect security than the traditional method. And the embedded capacity of the method gets better increase than the method based on the attributes of tags. So the method could be used to protect the content of a webpage and covert communication.

The Method

The Proposed Scheme. In this section, we will present the proposed information hiding scheme based on the slash of webpage tags. The process diagram of the scheme, which is composed of the information embedding process and the information extraction process, is shown in Fig. 1. The embedding process is used to hide the secret data in the cover webpage, while the extraction process is used to extract the secret data from the hidden webpage.

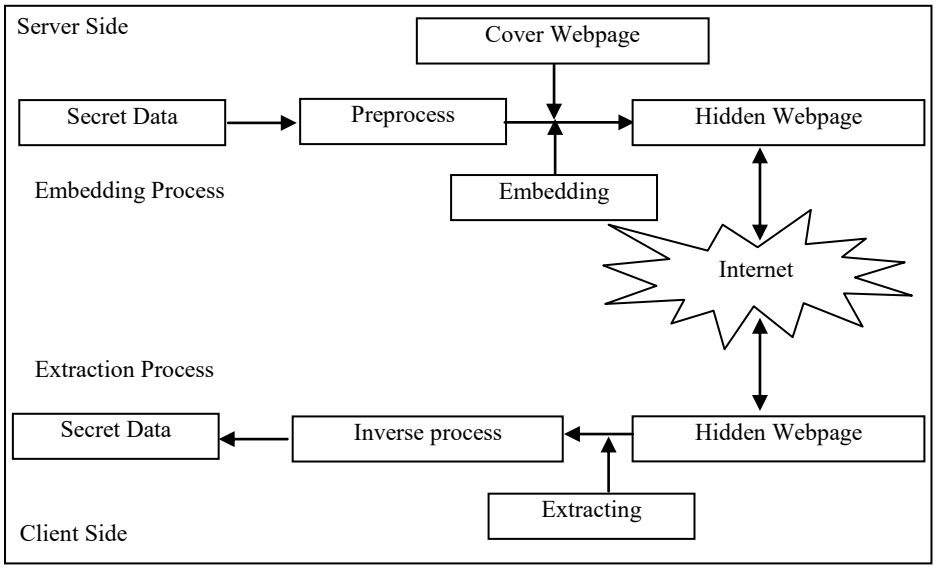

Fig 1. The Process Diagram of the Proposed Scheme

The Related Theoretics.

Definition 1. Let $T=<a_1, a_2, \ldots, a_n>$ be a tag with n attributes in HTML, where T is the name of the tag and a_i $(1\leq i\leq n)$, whose general form is "attribute name=attribute value"(short for name=value), is the i-th attribute of the tag. And let Ts be a single tag without end tags in HTML, meanwhile let Td be a double tag with starting and ending Tags. The starting component of any tag is the tag name and its attributes, if any. The corresponding ending tag is the tag name alone, preceded by a slash (/). Ending tags have no attributes.

Definition 2. Let $|W|$ be number of webpage tags, where W is a webpage. And let $|T_i|$ be a number of attributes of the i-th webpage tag, where T_i is the i-th tag in the webpage.

Definition 3. Let O be a object that is composed of a attribute and a value of the attribute in a tag, where O_i is the i-th object in the tag. For example, the "size=21px" is O_1 and the "color=green" is O_2 in the tag "".

Definition 4. Let T and T' be a pair of equal tag object, where T is a tag object without slash marks, T' is a tag object with a slash marks. For example, the T is "", the T' is "". That is $T \Leftrightarrow T'$.

By studying thoroughly, we found that the view results do not occur any change between the original webpages and the modified webpages using the equal attribute object in the browser.

Property 1. Equal attribute object has the identical function.

Rule 1. Embedded decision rule.

If a tag object T is a Ts tag or the starting component of any Td tag such that the "
" tag is a Ts tag and the "" tag is the starting component of Td tag " ", the T meets the insertion requirement; Otherwise, the T does not meet the insertion requirement.

Rule 2. Extraction decision rule.

If a tag object T' with the slash marks is a Ts tag or the starting component with the slash marks of any Td tag such that the "
" tag is a Ts tag and the "" tag is the starting component of Td tag " ",, the T' meets the extraction requirement; Otherwise, the T' does not meet the extraction requirement.

Rule 3. Embedded rule.

Step 1: Let i = 1, where $1 \le i \le \sum_{k=1}^{|W|} |T_k|$

Step 2: If the T_i meets the insertion requirement of Rule 1, then go to Step 4.

Step 3: Let i = i + 1. If the $i \le |W|$, then go to Step 2. Otherwise, go to Step 5.

Step 4: The T_i is modified to the T_i' , and let i = i + 1.If the $i \le |W|$, then go to Step 2.

Step 5: Finished.

Rule 4. Extraction rule.

Step 1: Let i = 1, where $1 \le i \le \sum_{k=1}^{|W|} |T_k|$

Step 2: If the T_i meets the extraction requirement of Rule 2, then go to Step 4.

Step 3: Let i = i + 1. If the $i \le |W|$, then go to Step 2. Otherwise, go to Step 5.

Step 4: We extract a secret information bit, and let i = i + 1.If the $i \le |W|$, then go to Step 2.

Step 5: Finished.

The Hiding Process. Let $W=\{ T_1, T_2, \ldots, T_n \}$ be a cover webpage, where T_i is a tag in the webpage. And let $M=\{ m_1, m_2, \ldots, m_n \}$ be the secret data bits to be embedded in the cover webpage. In order to increase the secrecy of the proposed scheme, we generate a chaotic sequence $L=\{ l_1, l_2, \ldots, l_n \}$, accompanied by a secret key to manipulate it, by the Logistic map system. Then the secret data bits is calculated by using the Eq.1.

$$S=M \oplus L=\{ s_1, s_2, \ldots, s_n\}=\{ m_1 \oplus l_1, m_2 \oplus l_2, \ldots, m_n \oplus l_n\} \quad (1)$$

We use the $S=\{ s_1, s_2, \ldots, s_n \}$ to determine whether the tag can be used to hide information or not according to the Rule 3.

The hiding process can be described as follows:

Step 1: Calculate the M form the secret data and generate the L by the Logistic map system and the secret key K, then Calculate the S from the M and the L.

Step 2: According to the Rule 1, check every tag object T_i of the webpage to determine whether the tag object T_i can be used to hide information or not. In our new method, if there is a equal tag object T_i' in the tag T_i , then the tag T_i is called embeddable tag.

Step 3: For the embeddable tag object, if the secret data bit of the S is 1, then replace the T with the T' for information hiding according to the value of the secret date bit. Otherwise, if the secret data bit is 0, the scheme retains the original tag object.

The payload capacity of the proposed scheme is given by Eq.2

$$\text{Capa} = \sum_{k=1}^{|W|} |T_k| \quad (2)$$

The Extraction Process. In this subsection, we shall describe the extraction process. The following extraction procedure is used to extract the embedded secret data. The extraction process can be described as follows:

Step 1: According to the Rule 2, check every tag object T_i of the webpage to determine whether the tag object T_i has been used to hide information or not. In our new method, if there exists the tag object T_i' , then the hidden secret data bit s_i is 1; otherwise, the hidden secret bit s_i is 0.

Step 2: Since the receiver owns the secret key used to generate the chaotic sequence L by the Logistic map system, the original secret data can be calculated by using the Eq.3.

$$M=S\oplus L=\{m_1, m_2, \ldots, m_n\}=\{ s_1\oplus l_1, s_2\oplus l_2, \ldots, s_n\oplus l_n\} \quad (3)$$

The Experiments

The experiments were carried out to evaluate the performance of the proposed information hiding scheme based on the slashs of tag in the webpage. The proposed scheme was tested on Win 7 personal computer with a Pentium IV 2.66GHz and 4G RAM. And six homepages of pop website were used as the cover webpage.

The Experimental Results. We have implemented the proposed information hiding method in the Visual C++ 6.0 environment. The experiment result shows that the view results did not occur any change between the original webpages and the modified webpages using the equivalent tag object in the browser. Fig. 2 and Fig.3 show the embedded secret data before and after webpage renderings. Fig. 4 and Fig. 5 show the source screenshots of the embedded secret data before and after webpage.

Fig. 2. The WebPage Rendering Before Embedded Secret Data

Fig. 3. The WebPage Rendering After Embedded Secret Data

```
<html> <body bgcolor=gray>
<table> <tr>
<td width=200 align="center"> <font color=red size=5> Test</font></td>
</tr> <tr>
<td width=260 align="right" height=50> Hiding Information</td>
</tr></table>
</body></html>
```

Fig. 4. The Source Screenshot Before Embedded Secret Data

```
<html> <body bgcolor=gray />
<table> <tr />
<td width=200 align="center"> <font color=red size=5 /> Test</font></td>
</tr> <tr />
<td width=260 align="right" height=50 /> Hiding Information</td>
</tr></table>
</body></html>
```

Fig. 5. The Source Screenshot After Embedded Secret Data “01010111”

At the same time, the homepage on the popular website has been tested for the maximum hidden amount of webpage. Table 1 shows the largest embedded capacity, which is called LEC for short, of the homepages on some popular websites which were visited on June 5, 2012. And the experimental results show that the method has good imperceptibility and perfect security than the method which was proposed in [6].

Table 1 The LEC of Homepages on Some Popular Websites

Homepage of Website	Method		
	LEC(b) of our method	*LEC(b) of method in [6]*	*LEC(b) of method in [7]*
www.163.com	2701	376	3984
www.yahoo.com	908	215	1332
www.microsoft.com	821	127	3242
www.sohu.com.cn	1826	624	4606
www.ebay.com	111	38	147
www.sina.com.cn	2753	1892	8274

The Performance Evaluation. Table 2 shows the performance parameters of our method and other six algorithms which are the invisible character, the changing uppercase and lowercase of tag, the order of attributes pair, the equal tag displacement and the repeat attribute of the tag algorithm.

The invisible character method which do not effect the normal show hide the secret information by adding spaces and tabs on the end of the line, but those who do not see characters at a glance when the source code of the hidden information page were selected simply. The changing uppercase and lowercase of tag method which do not effect the normal show too hide the secret information by using the tags characteristics of case-insensitive in the HTML norms, however the artificial alter of the tags is very easy found by observing the source code of the hidden secret information webpages, and then the hidden message was exposed. The changing order of attributes pair method which has strong anti-testing capability hide the secret data by changing the order of attributes pair in the webpage, but the extracting secret data needs the original database which generates additional transmission at the time of transporting. The equal tag displacement method in [6] which has strong anti-testing capability do not change the original file size after hiding secret information. Even viewing the page's source code can not determine whether hidden secret information in the webpage. But we can not hide any information in the page while all tags of page have one attribute at most. The equal tag attributes method in [7] do not almost change the original file size after hiding secret information ,and the equal tag attributes is not easy found by observing the source code of the hidden secret information webpages, but it has not strong anti-testing capability and the hidden secret information was not extracted correctly, if someone changed the order of the equal tag attributes. The repeat attribute of tag method in [8] changes the file size on bigger degree and the repeat attributes of the tag is very easy found by observing the source code of the hidden secret information webpages, and then the hidden secret information was exposed.

Our method does not change the display of the content and appearance of webpages after hiding secret information. The hidden secret information was not found by viewing the source code of the webpage. And then anyone can't change the order of the equal tag or else the webpage was not display properly. According to the above experiment result, our method has strong anti-testing capability, strong security capability, strong robustness capability, good imperceptibility and larger embedded capacity than other methods, such as the equal tag displacement method in [6] and the equal tag attributes method in [7].

Table 2 Performance Parameters of Five Algorithms

Method	Parameter				
	Imperceptibility	*Robustness*	*Change File Size*	*Against of Detection*	*Security*
Invisible Character	Good	Weak	Yes	Weak	Weak
Changing Case of Tag	Good	Weak	No	Weak	Weak
Order of Attributes Pair	Good	Strong	No	Strong	Strong
Equal Tag Displacement in [6]	Good	Strong	No	Strong	Strong
Equal tag Attribute Displacement in [7]	Good	Strong	Yes	Strong	Strong
Repeat Attribute of Tag in [8]	Good	Strong	Yes	Strong	Strong
Our Method	Good	Strong	Yes	Strong	Stronger

Conclusions and future works

Information hiding technology is a hot spot in information security, and is applied in the fields of digital multimedia copyright protection and secret communication. According to the analysis of the characteristics of browser in parsing HTML of the webpage and the little capacity available for information hided in webpage, a new efficient webpage information hiding method with equal tag has been proposed in this paper, which overcomes the shortcoming of the ability of imperceptibility and the ability of contradict with the machine filtration of traditional webpage information hiding algorithms and improves the embedded capacity of the other algorithm based on tag attributes. This method has good performances in invisibility and higher applied value as proved by the experiments. So we conclude that the proposed method is practical in many real applications.
The next work is to study how to improves the embedded capacity and security capability of the method by using the relative links of the webpages and multi-webpage embedment technology or other ones.

Acknowledgment

This work is supported by the Constructing Program of the Key Discipline in Huaihua University, by Scientific Research Fund of Hunan Provincial Education and by Scientific Research Fund of Huaihua University(HHUY2011-17, 201125).

References

[1] F. A. P. Petitcolas, R. J. Anderson, M. G. Kuhn, Information hiding-A survey, Proceedings of the IEEE, vol.3, 1999, pp. 1062–1078.

[2] C. John, Hiding Binary Data in HTML Documents, http://www.codeproject.com/csharp/steganodotnet13.asp, 2011-12-24.

[3] Q. J. Zhao, H. T. Lu, A PCA-based Watermarking Scheme for Tamper-proof of Web pages, Pattern Recognition, Elsevier Science, Oxford, ROYAUMEUNI, 2005, vol.38, no.8,pp.1321-1323.

[4] Q. J. Zhao, H. T. Lu, X. H. Jiang, Web page Watermarking for Tamper-proof, Journal of Shanghai Jiaotong University(Science), China, 2005, vol. 3, no.E-10,pp.280-284.

[5] L. Hu, X. G. You, Analysis of HTML information hiding technology, In Proc of CIHW2001, Xi'an: Xidian University Press, 2001, pp. 62-67.

[6] X. M. Sun, H. J. Huang, B. W. Wang, G. Sun, J. W. Huang, An Algorithm of Webpage Information Hiding Based on Equal Tag, Journal of Computer Research and Development, 2007,vol.44,no.5,pp.756-760.

[7] Y. J. Yang, Y. M. Yang, An Efficient Webpage Information Hiding Method Based on Tag Attributes, Fuzzy Systems and Knowledge Discovery, 2010 Seventh International Conference on,2010,vol.3,no.7,pp.1181-1184.

[8] J. G. Li, X.H. Ma, X. F. Shen, A Novel Scheme of Multiple Webpages Information Hiding Based on Repeating Tag Attributes, Computer Applications and Software, 2009, vol.26, no.8, pp. 62-64.