HTML Steganography using Relative links & Multi web-page Embedment

[1]Chintan Dhanani, [2]Krunal Panchal

[1]Mtech Scholar, [2]Lecturer
Department of computer engineering, Gujarat Technological University, Gujarat, India
chintan00014@gmail.com [1] chintan00014@gmail.com, [2] krunaljpanchal@gmail.com

***Abstract* – Cryptography, Steganography & Watermarking are three basic methods to protect data from unauthorized access. Cryptography & steganography are data hiding techniques which are used to send data securely while watermarking is used to give unique identity to data like image, audio & video. This paper concentrates on HTML steganography which uses HTML source code as cover text to hide data behind it. HTML steganography methods like use of null spaces, attribute order, attribute value enclosures, case of characters, hexadecimal code to hide data have some limitation like limited LEC(Largest Embedding Capacity) & law security. To remove this limitations better idea is to use relative links of HTML source code & Multi web-page embedment for embedding data in multiple web pages. That increases the LEC & security.**

***Keywords* – Steganography; LEC; Stego key; Stego cover; Carrier; Embedding; Extracting; Relative Links; Multi Web page Embedment**

I. INTRODUCTION TO HTML STEGANOGRAPHY

"Steganography" is a Greek word which means concealed writing or hidden writing. Steganography is the art and science of encoding hidden messages in such a way that no one except the sender and intended recipient, suspects the existence of the message. Steganography uses audio, video, text, Huffman code etc as carrier of information. HTML steganography is one part of text steganography which uses HTML web document as a carrier. To use HTML as a carrier has some benefits like large amount of cover documents available to hide data & by that decoding of that data by any unauthorized user is very difficult.

Basic model of steganography

Fig-1 gives the idea about steganography scheme in which first step is to embed original message in the carrier using any embedding technique then embedded message travel through the transmission media. At the receiver side receiver decodes the message which is the reverse process of embedding and gets the original message.

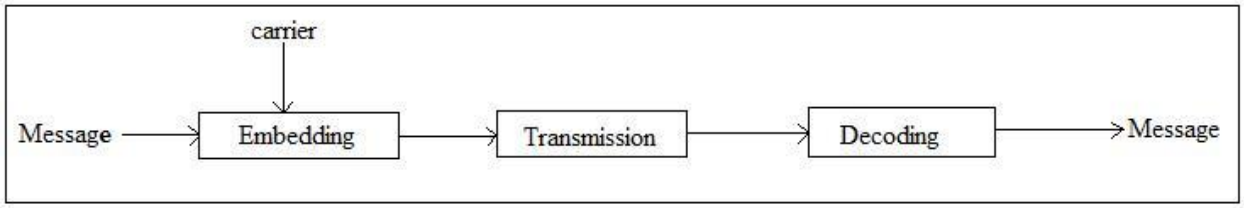

Fig 1 Basic model of steganography

HTML steganography Techniques

1) By changing order of Attributes[1] : HTML tag contains numerous amount of attributes & attribute order in the tag doesn't affect the output of the web pages. So we can hide data using attribute order.

Example:

<body **background="**image1.jpeg**" bgcolor="**#FFFFFF**">**……0
<body **bgcolor="**#FFFFFF**" background="**image1.jpeg**" >**.....1

As per above example if the sequence of attributes is (background,bgcolor) then it hides 0 and if sequence of attributes is (bgcolor,background) then it hides 1.

2) By using Null Space or White space or Invisible character [2]:

Example :

Stego key:

<tag>,</tag> or <tag/>……0
<tag >,</tag >, or <tag / >…..1

Stego data :

<customer ><name>James</name ><id >2345</id></customer> Embeded data :

101100

As per above example if tag contains the white space then it hides 1 & if tag doesn't contain the white tag then its hides the 0.

3) Modify the written state (case) of letters[4]:
Example:

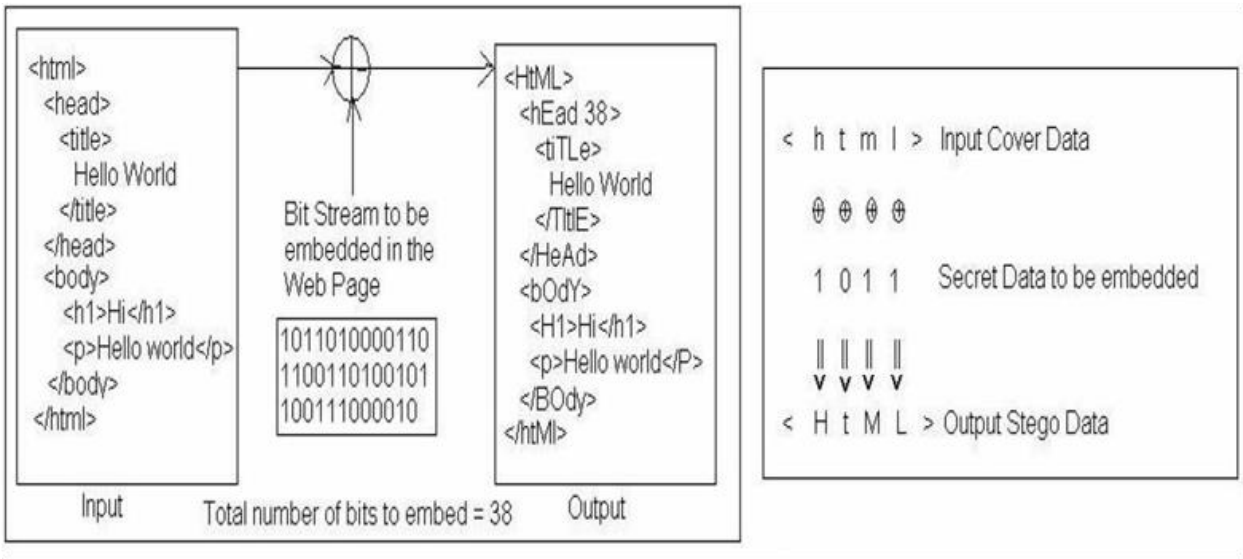

Fig 2 Hide data by changing case of letters

As shown in Fig 2 if the case of letter is converted from small to capital then it hides 1 & if case of letter remains same then it hides 0.

4) Color code or tag id replacement with Hexadecimal data[3] :

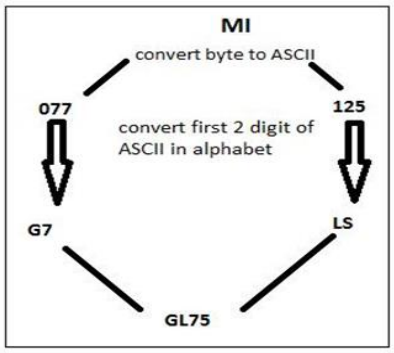

Fig 3 Character to Hexadecimal Conversion

Fig 3 gives idea about how message converted into Hexadecimal code & the hexadecimal form of message look like id attribute of tags or id of form elements. So we can place these hexadecimal code as value of id in HTML pages.

II METHODOLOGY ADAPTED [5]

The adapted methodology used in experiment uses attribute value enclosures to hide data. Attribute value can be enclosed with single inverted comma, double inverted comma or we can put value without comma. It doesn't affect the output of HTML page. This method can hide number of bits same as number of attributes. Fig. 4 and Fig.5 show the embedded secret data before and after web page rendering. Fig. 6 and Fig. 7 show the HTML source code of the embedded secret data before and after embedding.

Fig 4. HTML page output before embedding data

Fig 5. HTML page output after embedding 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 6. Source code before embedding 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 7. Source code after embedding data '10010110'

Limitations of adapted methodology

Method can hide maximum bits as same as number of attributes so if the message has more bits then it is not possible to embed whole message & other problem is whole message is embedded in one page can be easily detected by comparing original page with stego-page. So in this method LEC is law that should be increased & security should be improved.

III.PROPOSED SCHEME & ALGORITHM

As we see HTML steganography technique has some limitations like limited LEC & law security. These problems can be solved by using Relative Links of web pages & multi web page embedment Technology[5].

A) Brief Idea about Proposed Scheme

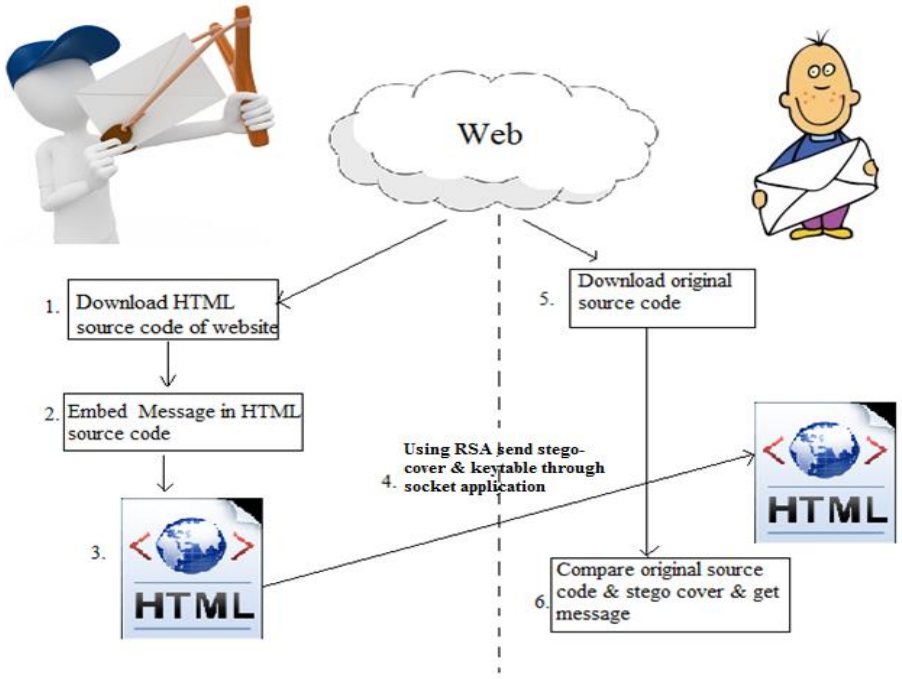

Fig 8 Brief Idea About Proposed Scheme

Step 1: Sender Downloads whole source code of any web site & generate table having columns (page_name, page_link, page_status,page_visiting_no). The default page_status value is set as 'unvisited' for all downloaded web pages.
Step 2: Embed data in multiple web pages using relative links & multi web page embedment technology & generate stego cover having data embedded in it.
Step 3: Sender sends stego cover to Receiver using RSA.
Step 4: Receiver receives stego-cover in encrypted form.
Step 5: Decrypt the stego-cover & compare stego-cover with original source code & get the message hidden in stego-cover.

B) Data Embedding Algorithm
Step 1: count = number of href tags in the page
Step 2: median = int(count/2)
Step 3: Hide the data in the page until median href tag encountered.(using changing enclosure method).
Step 4: Then transfer control to page having relative link= median href tag.
Step 5: if page is already visited then find the page from the key table having page_status=unvisited & mark it as visited.
Go to step 1.
Follow these 5 steps until data hiding is over.

C) Data Extracting Algorithm
Step 1: find the page in key table having visiting no='1'.
Step 2: count href tag in the page & take median = int(count/2)
Step 3: Extract the data in the page until median href tag encountered.
Step 4: visiting no ++;
Step 5: Go to step 2
Follow steps 2 to 5 until entire message extracted.

IV. EXPERIMENTAL RESULTS & PERFORMANCE EVALUATION

Algorithm uses relative links (href tags) of web pages to transfer from one web page to another web page. It divides message in more than one parts & embed it in different pages. So now we have available more than one pages to embed the single message that increases LEC & improve security because message cannot be extracted without the keytable containing the visiting number of pages.

A) Experimental Results

Results gives Comparison of two methods. One method uses attribute value enclosure method & our method uses attribute value enclosure method with relative links & Multi web page embedment. We have tested results on the source code of the site www.atmansol.com

1) Results for Attribute value Enclosure Method

HTML pages of site www.atmansol.com	**LEC(Largest Embedding Capacity)**
aboutus.html	412 bits

automation.html	444 bits
career.html	405 bits
consulting.html	428 bits
contactus.html	517 bits
custom_built_applications.html	442 bits
field.inspection.html	445 bits
index.html	623 bits
instrumentation_engineering.html	441 bits
manufacturing_solution.html	427 bits
panels.html	646 bits
privacy_policy.html	424 bits
project_management.html	439 bits
services.html	501 bits
sitemap.html	562 bits
support.html	452 bits
termofuse.html	411 bits

Table 1: Results of changing attribute value enclosure method

For this technique if we take the web page from the site www.atmansol.com that can have maximum LEC then the page panels.html have highest LEC. So using this method we can hide maximum 646 bits for this site.

2) Results for method which using relative link & Multi web page Embedment Technology

HTML pages of site www.atmansol.com	**LEC(Largest Embedding Capacity)**
aboutus.html	248 bits
Automation.html	295 bits
Career.html	241 bits
Consulting.html	264 bits
Contactus.html	353 bits
Custom_built_applications.html	258 bits
Field.inspection.html	265 bits
Index.html	368 bits
Instrumentation_engineering.html	247 bits
Manufacturing_solution.html	236 bits
Panels.html	359 bits
Privacy_policy.html	265 bits
Project_management.html	239 bits
Services.html	281 bits
Sitemap.html	311 bits
Support.html	263 bits
Termofuse.html	235 bits
	TOTAL : 4728 bits

Table 2: Results of method using relative link & Multi web page embedment

Use of relative link can hide data in any web page of the site & Multi web page embedment can hide data in more then one pages of the site. So here we have total number of bits that can be hidden in the web site www.atmansol.com is 4728. So LEC is dramatically increased as we want to improve it.

B) Performance Evaluation

Performance evaluation can be done on the parameters LEC & security

Method/Parameter	**LEC**	**Security**
Using Null space or white space	Law	Weak
Modify the written state(case)	High	Weak
Changing Attribute Order	Law	Strong
Tag displacement	Law	Strong
Changing Attribute value enclosures	Medium	Average
Color code or tag id replacement with HEX code	Medium	Average
Our Method	High	Strong

By getting the results of all the HTML steganography methods & considering two parameters LEC & security table has been generated. That shows the method using relative link & multi web page embedment gives high LEC & strong security in compare to other methods.

VI. CONCLUSION

HTML steganography is new era of hiding data & it gives more feasibility to hide data because there is huge number of pages available on the internet & data hidden behind HTML pages is less suspicious. In this paper we have use relative links & multi web page embedment technology to increase LEC & security of data that gives freedom to send larges message securely over the internet.

REFERENCES

[1] Mohit Garg, "A Novel Text steganogrpahy Technique Based on HTML Document", International Journal of advanced Science and Technology Vol. 35, October 2011.
[2] Shingo Inoue, Ichiro Murase, Osamu Takizawa, Tsutomu Matsumoto, Hiroshi Nakagawa, "A Proposal on information Hiding Mehods Using XML"
[3] Mohammad Shirali Shahreza, "A New Method for Steganogrpahy in HTML Files", Advance in computer, Information and System Science & Engineering , 247-251, 2006 Springer.
[4] Xin-Guang Sui, Hui Luo, "A new Steganography method based on Hypertext", IEEE-2004.
[5] Yujun Yang, Yimei Yang, "An Efficient webpage Information Hiding Method Based on tag Attributes", IEEE-2010.
[6] http://www.atmansol.com