

A Modified Approach to Text Steganography using HyperText Markup Language

Susmita Mahato¹, Dilip Kumar Yadav², Danish Ali Khan³

Department of Computer Applications

National Institute of Technology

Jamshedpur, India-831014

sus_ps@rediffmail.com¹, dkyadav1@gmail.com², dakhan_nitjsr@sify.com³

Abstract—Steganography is the tactics of silent communication, where one person communicates with others through cover medium so that intermediary does not has suspicion about the hidden information. In this paper we propose a modified approach for text steganography based on HTML tags and attributes. As HTML is rich in tags and its attributes, easily communicated in the internet, and the source code is checked rarely by anybody it can be used intelligently to perform text steganography. By hiding the secret data inside the source code of HTML, text steganography can be easily achieved. In this approach the relation between two consecutive attributes is considered for hiding secret data. To hide '1' two consecutive attributes of same tag are taken, and to hide '0' two consecutive attributes of different tags are taken. The secret data hiding and revealing technique is presented. The proposed approach has been explained using java programming language.

Keywords—steganography; cryptography; Hypertext Markup Language; attributes

I. INTRODUCTION

In this digital age, communication is made easy with the availability of the internet. In last few years the volume of data packets has increased many folds and data traffic are ranging from kilobytes to terabytes. With increase in data traffic through various types of transceivers over varied communication media, the security of the vital data is of great concern. Cryptography and steganography are used for securing data over communication channel. Cryptography focuses on making the message contents secret, whereas steganography's focus is to keep the presence of the message itself secret.

Researches on Cryptography are going on from ancient era. Cryptography technique was started from 1900 BC when a scribe in Egypt first used a derivation of the standard hieroglyphics of the day to communicate [1]. In modern age communication both fields, Cryptography and Steganography is getting developed hand in hand, however in cryptography a lot of research has been done, comparatively steganography still has to go long way.

Steganography is a technique to hide secret data in some cover media as text, image, video, audio etc. Among these media text Steganography is very much difficult, as there are less abundant bits to store the secret bit information.

HTML (Hypertext Markup Language) was developed by scientist Tim Berners-Lee in 1990. It is the hidden code that makes us able to communicate with others on the World Wide Web. Here tags are used with the text in order

to create the structure, which tells the browser how to display the text or graphics in the document but do not show themselves. As HTML is very simple, can be composed and edited on any computer as windows, Unix etc. It is very rich in tags and its related attributes, so we can hide very large amount of data on it. It is easily communicated in the internet, and the source code is checked rarely by anybody. The size of HTML file is also very less compared to image, video or audio file. It can run on every browser. So HTML source code can be used for text steganography very conveniently.

This paper explains the development in the field of text steganography and makes an attempt to modify the conventional approach of text steganography using HTML in an efficient way.

The rest of the paper is organized as follows: Section II discusses basic concepts; Section III shows related works. Section IV describes the proposed technique. Section V demonstrates working of our approach. Section VI gives experimental results. Conclusion is given in Section VII.

II. BASIC CONCEPTS

Commonly two types of information hiding techniques are used to hide and secure data transmission in communication media.

A. Cryptography

Cryptography is a Greek word mean hidden writing. It is a science of protecting secret message by converting plain text into cipher text by some encryption method or algorithm. Thus message looks like meaningless data. Figure.1 shows a typical cryptography mechanism.

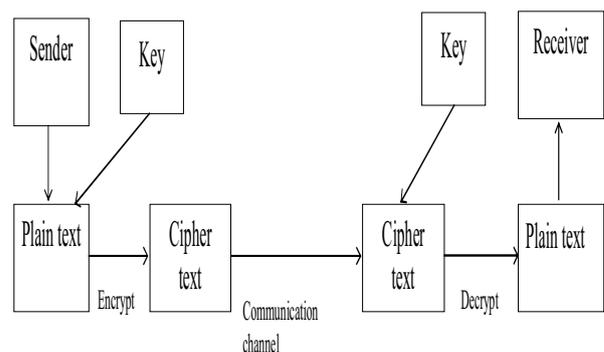


Figure 1. Cryptography mechanism

B. Steganography

Steganography has been used from writing of Herodotus, as Histiaeus communicated to his son-in-law in Greece to revolt; by shaving the head of his most trusted slave and tattooing it with a message which disappeared after the hair had grown-up back. Thus the slave delivered the message his way from Peria to Greece [2].

Steganography is a Greek word. Here Stegos means cover and Grafia means writing, i.e. covered writing. Steganography means hiding secret data by some another data, that can be text, picture, audio or video data. Steganography protects data by hiding secret message in some cover media, thus conceal their very existence. Figure.2 shows a typical Steganography mechanism. Steganography may be classified according to the types of cover media used for hiding secret data.

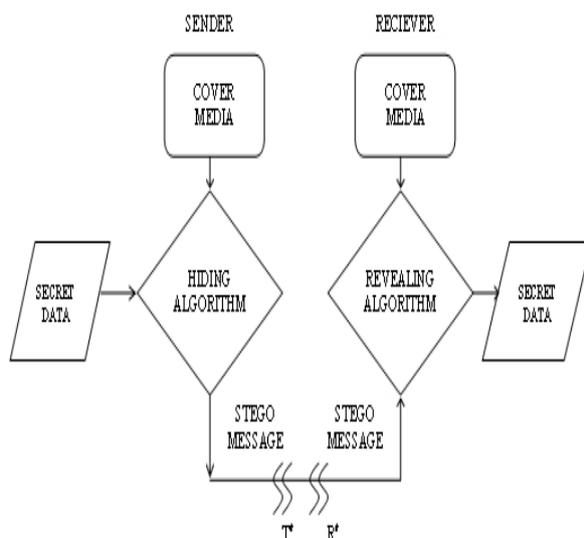


Figure 2. Steganography mechanism

III. RELATED WORKS

Different researchers proposed different text steganography techniques. S. H. Low, N. F. Maxemchuk, J.T. Brassil, and L.O'Gorman proposed a technique for text steganography achieved by line shifting method [3]. Here shifting by some degrees as 1/300 inch up or down to the lines of text are done and thus different unique shapes of the text are used for hiding the information. In word shifting method, hiding of information is done by shifting the words more or less i.e. changing the gap between the words.

Moreland proposed a text steganography technique based on using selected characters from word [4]. As example the first letter of all paragraphs can be used to conceal the secret message. By placing all selected characters together we get the complete secret message.

D. Huang and H. Yan proposed a text steganography technique based on putting extra white-spaces in between the text [5].

Shirali-Shahreza proposed a text steganography technique based on different spellings of words. In British and

American English, some words have different spelling as color and colour [6].

Mohit Garg proposed a technique of text steganography based on HTML document where primary and secondary attributes were taken and based on the requirement of '0' or '1' position alteration was done from primary to secondary [7].

Aasma Ghani Memon, Sumbul Khawaja and Asadullah Shah proposed a technique of text steganography based on XML document [8].

Herman Kabetta, B. Yudi Dwiandiyanta and Suyoto proposed a technique of text steganography where information was hidden in Cascading Style Sheet [9].

IV. THE PROPOSED TECHNIQUE

Optimum hiding without compressing with any loss in data, or their meaning, is of utmost concern. Our approach is based on HTML. The proposed technique is a modified technique in which cover medium HTML page is used to conceal the secret message.

Several tags with several attributes are used for making an HTML page. Our proposed technique is based on the tags and attributes. We have coded the proposed technique in Java language. It has following components:-

- 1) Tag's attribute table
- 2) Key to generate more html attributes to improve security
- 3) Method to hide secret message in HTML page attributes
- 4) Method to reveal secret message by checking pair present or not by mapping with the tag's attribute table.

A. Algorithm for hiding secret message in HTML file

Step 1:- Take the secret message

Step 2:-Put HTML tags attributes from TABLE 1, as for first 1 put consecutive pair from same tag, and for 0 from two different tags. And follow the same procedure for the next bit by comparing with the current one.

Step 3:-Take a key, Match with the attribute list obtained from step 2 from left to right, in whichever positions 1 is present in the key bit list, take the attribute else if 0 is present put any other attributes, hence the list of attributes for making HTML page is increased. Key should have one more number of 1's same as the length of the secret data.

Step 4:- Make the HTML page using attribute list with their respective tags.

Step 5:- Put values to the attributes.

Step 6:- Send it to receiver.

B. Algorithm for extracting secret message from HTML file

Step 1:- Make table of all attributes present in the HTML page by scanning HTML page's source code. Extract each tag's attribute list and combine them together.

Step 2:-Put key bit with each element, then make another list of attributes by discarding the attributes in which 0 is present in the key bit.

Step 3:- Map the attribute with tags attribute's table file. If consecutive attribute pairs from the same tag are available then put 1 else put 0, and hence we get the data.

V. WORKING OF OUR APPROACH

A. Hiding secret message

The proposed hiding process is illustrated in Figure. 3.

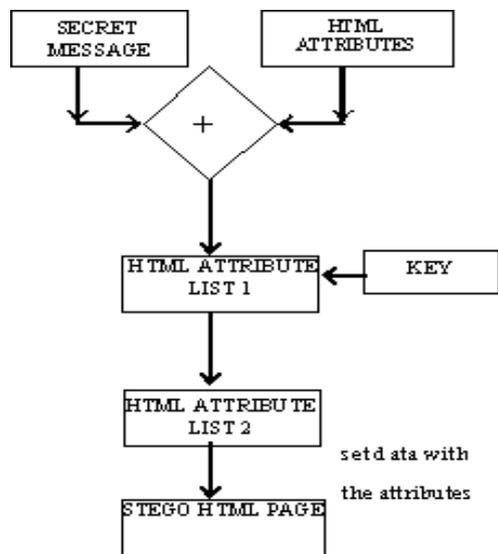


Figure 3. Hiding process

- 1) The secret message is: - 11100
- 2) Put html tags attributes. As for 1, put consecutive pair attributes from Tag-Attributes Table (Table I).
bgcolor background alink vlink size src
- 3) Take a key as 1650 (11001110010) as it should contain same number of 1's as the number of total attributes obtained from (2) i.e. (total number of bits present in secret message) +1.
- 4) Match with the secret message from left to right and put 0 in between
bgcolor background 0 0 alink vlink size 0 0 src 0

Put any tag attribute in place of 0

```
bgcolor background A B alink vlink size C D src E -
- - - -
bgcolor background text link alink vlink size face color
src alt align border height hspace vspace
```

5) Make the html page.

```
<Html>
<Head>
<Title>example</Title>
</Head>
<Body bgcolor= "color" background="filename"
text="color" link="filename" alink="filename"
vlink="filename">
<Font size="value" face="name" color="color">
I am showing an example
</Font>
<Img src= "url" alt ="text" align="direction"
border="number" height="pixels" hspace="pixels"
vspace="pixels">
</Body>
</Html>
```

B. Revealing secret message

The proposed revealing process is illustrated in Figure. 4.

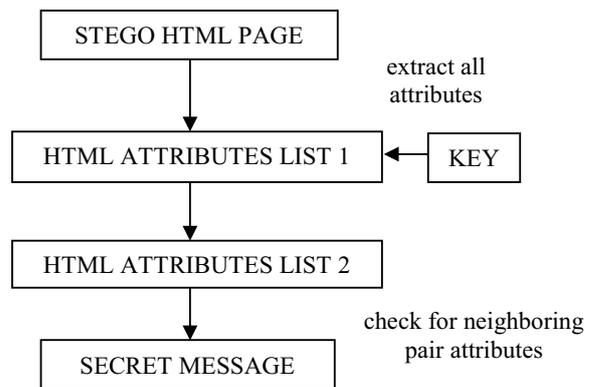


Figure 4. Revealing process

Example:-
The covered HTML document is:-

```
<Html>
<Head>
<Title>example</Title>
</Head>
<Body bgcolor= "color" background="filename"
text="color" link="filename" alink="filename"
vlink="filename">
<Font size="value" face="name" color="color">
I am showing an example
</Font>
<Img src= "url" alt ="text" align="direction"
border="number" height="pixels" hspace="pixels"
vspace="pixels">
</Body>
</Html>
```

TABLE I. TAG-ATTRIBUTES TABLE

Tag	Attributes
BODY	BGCOLOR, BACKGROUND, TEXT, LINK, ALINK, VLINK, ONLOAD, ONUNLOAD, CLASS, DIR, ID
FONT	SIZE, FACE, COLOR
IMG	SRC, ALT, ALIGN, BORDER, HEIGHT, WIDTH, HSPACE, VSPACE, USEMAP, ALIGN, LOWSRC
TABLE	BORDER, CELSPACING, CELLPADDING, HEIGHT, WIDTH, BGCOLOR
TH	ALIGN, VALIGN, BGCOLOR, WIDTH, ROWSPAN, COLSPAN
TD	COLSPAN, ROWSPAN, ALIGN, VALIGN, BGCOLOR, WIDTH
TR	ALIGN, VALIGN, BGCOLOR, WIDTH
FRAME	SRC, NAME, MARGIN WIDTH, MARGIN HEIGHT, SCROLLING
FORM	ACTION, METHOD, ACCEPT, ENCTYPE
INPUT	TYPE
SELECT	NAME
CAPTION	ALIGN
H1	ALIGN
H2	ALIGN
H3	ALIGN
A	CHARSET, HREF, REV, SHAPE, TARGET
HR	ALIGN, NOSHADE, SIZE, WIDTH
OL	TYPE, START
P	ALIGN
UL	TYPE

1) Extract all the attributes with their tags, we get a list:-

bgcolor background text link alink vlink size face color src alt align border height hspace vspace

2) Take keys binary form, and match with the list, if 1 is paired to an attribute, take it, else discard it.

Just think key is- 1650 (11001110010)

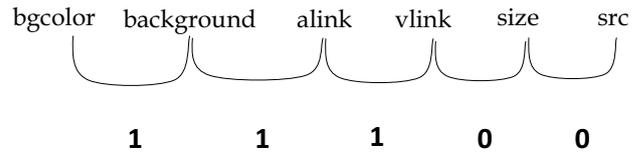
bgcolor background text link alink vlink size
 1 1 0 0 1 1 1
 face color src alt align border height hspace vspace
 0 0 1 0

The attributes not getting any pair, discard.

3) So, we get a new list as:-

bgcolor background alink vlink size src

4) Next check for consequent pair's presence in Tag-Attributes Table (Table I).
 If it is from same Tag then put 1, else put 0.



5) Thus the secret message is: - 11100.

VI. EXPERIMENTAL RESULTS

Here mainly two phases present the result of the proposed method of text steganography using Hypertext Markup Language. The proposed system is simulated using JAVA programming language. The following are the screen shots of the two phases for hiding and revealing the secret message, shown in the Figure 5, 6, 7, 8. This algorithm satisfies both the security and hiding capacity necessities. The key plays an important role in making the algorithm more secure. If key is changed by increasing or decreasing the number of '1' in key, the secret message cannot be extracted and if the stego HTML source code is changed by adding new attribute or deleting existing attribute then also the secret message is not extracted by the desired key.

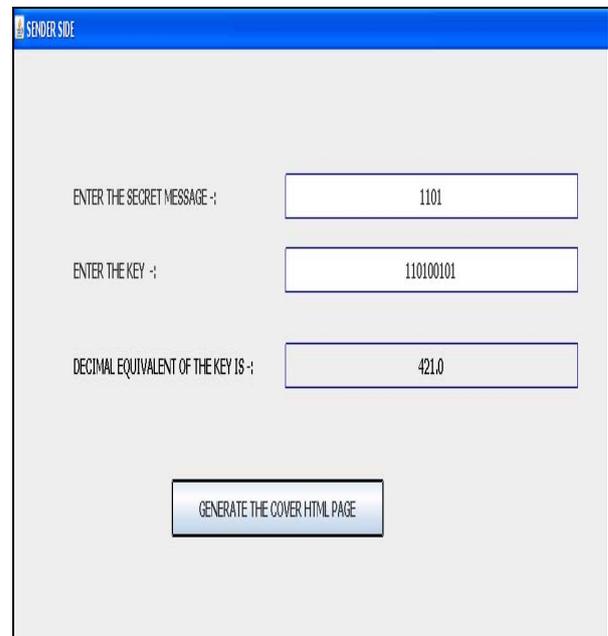


Figure 5. Generating stego HTML page

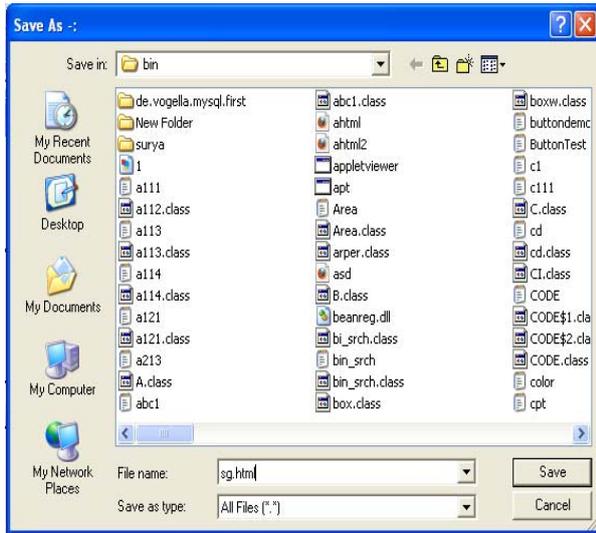


Figure 6. Saving generated Html page



Figure 7. The generated HTML page

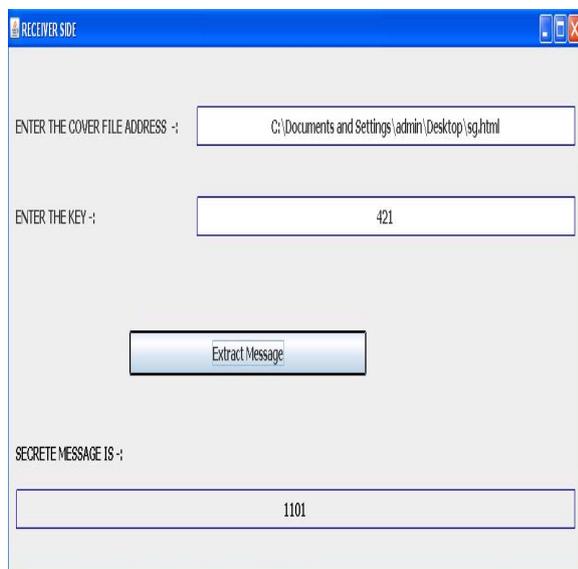


Figure 8. Extracting the secret message from HTML page

VII. CONCLUSION

In this paper text steganography technique for information hiding is presented. The paper proposes a modified approach to provide data security using HTML tags and attributes. As HTML is used for designing web pages, the secret data is hidden behind the attribute list of HTML tags, and hence web page is formed. This can be easily transmitted through the Internet to the destination. In the destination the receiver can view the source code of the web page, extract the attributes and hence following the revealing process the secret message is obtained. The proposed algorithm can also achieve high embedding capacity. Future research can be done to decrease the time complexity of the algorithm. Although this paper is based on HTML documents, it can be expanded to other document formats such as XML, CSS etc.

REFERENCES

- [1] A. Joseph Raphael, Dr. V. Sundaram, "Cryptography and Steganography – A Survey", International Journal on Computer Technology and Applications, vol. 2 (3), ISSN: 2229-6093, pp.626 – 630, 2010.
- [2] Herodotus: The Histories Penguin Books, London, Translated by Aubrey de Selincourt, 1996.
- [3] S. H. Low, N. F. Maxemchuk, J.T. Brassil, and L. O'Gorman, "Document marking and identification using both line and word shifting", Proceedings of the Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '95), vol. 2, pp.853 – 860, 2-6 April 1995.
- [4] T. Moerland, "Steganography and Steganalysis", www.liacs.nl/home/tmoerlan/privtech.pdf, 15 May 2003.
- [5] D. Huang, and H. Yan, "Inter word distance changes represented by Sine Waves for watermarking text images", IEEE Transactions on Circuits and Systems for Video Technology, vol. 11(12), pp.1237–1245, December 2001.
- [6] M. Shirali-Shahreza, "Text steganography by changing words spelling," Proceedings of the 10th International Conference on Advanced Communication technology, vol. 3, pp.1912–1913, 17-20 February 2008.
- [7] Mohit Garg "A novel text steganography technique based on Html documents" International Journal of Advanced Science and Technology vol. 35, pp. 129–138, October 2011.
- [8] Aasma Ghani Memon, Sumbul Khawaja and Asadullah Shah (2008) "Steganography: A New Horizon for safe communication through XML", Journal of Theoretical and Applied Information Technology, pp. 187–202, 2008.
- [9] Herman Kabetta, B Yudi Dwiandiyanta, Suyoto, "Information hiding in CSS: A secure scheme Text-Steganography using Public Key Cryptosystem", International Journal on Cryptography and Information Security (IJCIS), vol. 1(1), pp.13–22, December 2011.