Special Symbols Reordering HTML-Steganographic Method in Application to Wikipedia

I.V. Minin[a)] and O. V. Minin

National Research Tomsk Polytechnic University, Tomsk, Russia

[a)] *Corresponding author: prof.minin@gmail.com*

Abstract. Nowadays the Internet is a highly developed communication system. Every day the number of the users of the Wikipedia increases for thousands of people. Wiki database is open for everybody: either a law-abiding and decent person or a quite opposite one. This leads to such problems of the defense of information as copyright and transmission of hidden messages via World Wide Web through open communication channels. This article covers methods, which can help to solve these problems with usage of HTML based Wiki adapted steganographic algorithms.

INTRODUCTION

The wold “Steganography” was probably taken from a work by Johannes Trithemus, a German Benedictine abbot and a polymath, who in 1499 prepare and published in1606 the work entitled as “Steganographia”, where he developed his so-called "Ave-Maria-Cipher" that can hide information in a Latin praise of God: for example "*Auctor Sapientissimus Conseruans Angelica Deferat Nobis Charitas Potentissimi Creatoris*" contains the concealed word *VICIPEDIA* [1].

Historically the word “Steganography” comes from the Greek “covered writing” [2, 3]. For example, in around 440 BC, Herodotus [2] writes about Histæus, who shaved the head of his favorite slave, tattooed a message on his scalp, and waited for the hair to regrow, obscuring the message from guards – thus he was the first who being held captive and wanted to send a message without being detected.

Now the Internet is a public access open network which provides many benefits. But it is facilitating also the provision of an illicit communication channel whereby hidden messages can be received and sent by steganography, the process of hiding messages that the cover message remains unchanged at least to an unsuspecting observer. For example, according to UStoday, the usage of steganography by terrorists was suspected during the lead up to the 9/11 attack in the USA in 2001 [4].

HTML STEGANOGRAPHY

Usually webpages are written using Hyper Text Markup Language (HTML). The most common HTML steganography can be achieved in a variety of ways, for example

— Invisible characters Steganography [5]

— Attribute Permutation Steganography [6,7]

— Tag Letters Case Switching Steganography [8,9].

A number of steganographic techniques for hiding data in a cover were discussed in [10]. Sudeep Ghosh in [11] has shown how HTML TAGS can be manipulated to represent hidden bit ‘1’ or ‘0’. For example, in HTML documents, spaces are ignored as are carriage returns. The first space is accepted, but additional space(s) are explicitly ignored by the browser. Indeed Barilnik and Minins [12] has explored this and offer to used it to hide information in source code, HTML pages or anywhere [13]. The authors of Ref. [12] elaborates that software programs like ‘Steganos

for Windows' uses gaps i.e. space and horizontal tab at the end of each line, to represent binary bits ('0' and '1') of a secret message. The above example indicates hiding of secret bits '10011011' as per analogy explained (Table 1). As it was mentioned in [12] this can be useful for hiding a signatures for hidden data or copyright. It was also noted [12] that opening a HTML document in Microsoft Word and enabling the formatting marks, spaces are easily determinate.

TABLE 1. Windows' space as a secret bit.

Text	Code (secret bits)
<html>( )->->( )->	1001
<head>( )->( )( )->	1011

Here '->' denotes horizontal tab and ( ) represents space symbol.

WIKI STEGANOGRAPHY

Wikipedia is a multilingual online encyclopedia created and maintained as an open collaboration project by a community of volunteer editors and is a HTML-based publication database [14]. Let's consider some examples. Bits of hidden information are introduced in a form of unprintable symbols. To simplify the problem, below we will use only 3 special characters: HTML space " " ( ) (ASCII 32 (20)); HTML vertical bar "|" (|) (ASCII 124 (7C)) and HTML underscore "_" (_) (ASCII 95 (5F)) [15]. Each byte of hidden information is transformed into a succession of these symbols where each symbol corresponds with a bit of hidden byte [12].

EXAMPLE 1. INTERNAL LINKS (WIKILINKS)

The main idea is that the reordering of special wiki-symbols will not be removing or adding any hidden information from the file. Therefore standard techniques of HTML stego-analysis, which look for abnormal distribution of bits or/and text, should fail [16-18].

Underscores in the wiki-page title may be represented as spaces, this property can be used to encode hidden information. But using underscores in links will make them visible in the page text, therefore, we will use the following construction, shown in the first column of Table 2. The key of the assumptions behind this consideration is that special symbols ordering has a skewed distribution across all the wiki pages on the Internet. Based on the analysis of wiki-pages in the Internet, it can be concluded, that there is a real possibility a different programmers might arrange the special symbols in a random order because of wiki-pages directly managed and edited collaboratively by its own wiki-audience [14]. Let's look at some coding examples based on three Wikipedia pages [19-21].

TABLE 2. Underscore encoding character.

What you type	How it appears	Hidden codes
[[Vladilen F. Minin|V.F.Minin]]	V.F.Minin	No
[[Vladilen_F._Minin|V.F.Minin]]	V.F.Minin	11
[[Igor_V. Minin|I.V.Minin]]	I.V.Minin	10
[[Oleg V._Minin|O.V.Minin]]	O.V.Minin	01

Here: space symbol equal to "0" and underscore equal to "1".
It is important to note that the size of the file do not modified.

EXAMPLE 2. REFERENCE CITATION

Let's consider this possibility as an example of citing literature. To cite a professional or scientific journal in wiki-pages used the construction. In Wikipedia a full reference included the author's name, journal name, date, volume, issue, pages, etc. But special characters, for example, a space and an underline, in some cases will not be shown on the web page, which can be used to hide information. Moreover, it is essential that changing the order of this information, the browser will show it on web page in a standard form:

a) Most commonly used parameters in horizontal format

{{cite journal | *last*=Minin | *first*=Vladilen F. | last2=Serrano | first2=Daniel | last3=Minin | first3=Oleg V. | last4=Uris | first4=Antonio | last5=Minin | first5=Igor V. | year=2020 }}

b) Reordered

{{cite journal | *last5=Minin* | *first5=Igor V.* | last4=Uris | first4=Antonio | last=Minin | first=Vladilen F. | last2=Serrano | first2=Daniel | last3=Minin | first3=Oleg V. | year=2020 }}

c) Paddling

{{cite journal | *last4=Uris* | *year=2020* | *first=Vladilen F.* | *first2=Daniel* | *last5=Minin* | *first4=Antonio* | *last3=Minin* | *first5=Igor V.* |*last=Minin* | *last2=Serrano* | *first3=Oleg V.* }}

d) Combination

We may unite methods describe in Example 1 and Example 2 as follows:

{{cite journal | *last5=[[Igor V. Minin|Minin]]* | *first5=Igor V.* | last4=Uris | first4=Antonio | last=Minin | first=Vladilen F. | last2=Serrano | first2=Daniel | last3=Minin | first3=Oleg V. | year=2020 }}

In all these cases, it appears as: "Minin, Vladilen F.; Serrano, Daniel; Minin, Oleg V.; Uris, Antonio; Minin, Igor V. (2020)".

EXAMPLE 3

If the requirement, that the file size should not be changed, does not apply, then there are several possibilities for placing encoded information (Table 3).

TABLE 3. Vertical bar coding.

What you type:	**Hidden code**	
{{cite journal | last=Minin | first=Vladilen F.| last2=Serrano | …}}	no	
{{cite journal | last=Minin ||| first=Vladilen F.|| last2=Serrano | …}}	11	1
{{cite journal | last=Minin | | | first=Vladilen F. | | last2=Serrano | …}}	0101	01

Here: space equal to "0" and vertical bar equal to "1".

Again, in all cases it appears as: **"**Minin, Vladilen F.; Serrano,…".

These features provide significant opportunities for hiding the necessary information.

PSEUDO - HIDDING OF SECRET MESSAGES

The lifetime of messages in wiki-texts is limited by the time the users edit the text. And after editing, for the "normal" user, the message is erased (destroyed). But using the "Revision History" or "View history" mode with "compare selected revisions" option, you can see the initial text of the message, which remains in the archive of the wiki-page.

CONCLUSION

These articles dealt on the principles of hiding information in Wiki-pages. Different HTML-based steganographic methods in application to Wikipedia were offered and studied. The insight into the steganographic wiki-based principles will definitely guide us to improve its applications and to identify new information hiding detection methods. The results may be used to developed wiki-based copyright protection system similar to HTML system [22-23]. Proposed methodology allow embedding invisible watermark into Wiki-pages using the structural features of HTML-based language and this can be extended for other markup languages like SGML and XML.

ACKNOWLEDGEMENTS

This work was partially supported by TPU development program.

The paper was presented at Modern Approaches in Engineering and Natural Sciences: MAENS-2021 conference, Sept.15, Tver, Russia (2021)

REFERENCES

1. J. Reeds, Cryptologia **22(4)**, pp. 191–317 (1998).
2. F. A. P. Petitcolas, R. J. Anderson and M. G. Kuhn, "Information Hiding: A survey," in *Proceedings of the IEEE* (1999), pp. 1062–78.
3. Petitcolas and A. P. Fabien, *Techniques for Steganography and Digital Watermarking* (Artech House, Boston, 2000), pp. 1-14.
4. J. Kelley, *Terrorist instructions hidden online* (2001).
5. WbStego4open, http://www.wbstego.wbailer.com/.
6. H. Huang, S. Zhong and X. Sun, "An algorithm of webpage information hiding based on attributes permutation," in *4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (2008), pp. 257-260.
7. D. Shen and H. Zhao, "A novel scheme of webpage information hiding based on attributes," in *IEEE International Conference on Information Theory and Information Security (ICITIS)* (2010), pp. 1147-1150.
8. Y. Shen, J. of Wuhan University **50**, pp. 217-220 (2004).
9. X.-G. Sui and H. Luo, "A new steganography method based on hypertext," in *Radio Science Conference* (2004), pp. 181-184.
10. W. Bender, D. Gruhl, N. Morimoto, and A. Lu, *IBM Systems J.* **35(3&4)**, pp. 313-336 (1996).
11. S. Ghosh, *StegHTML: A message hiding mechanism in HTML tags* (2007).
12. S. S. Barilnik, I. V. Minin and O. V. Minin, *Adaptation of Text Steganographic Algorithm for HTML* (Novosibirsk State Technical University, 2007).
13. K. Bennett, *Linguistic Steganography: Survey, Analysis and Robustness Concerns for Hiding Information in Text* (Center for Education and Research in Information Assurance and Security West Lafayette, 2004).
14. Wikipedia, https://en.wikipedia.org/.
15. HTML Codes Table, https://ascii.cl/.
16. M. S. Shahreza, *A New Method for Steganography in HTML Files. Advances in Computer, Information, and Systems Sciences, and Engineering* (Springer, Dordrecht, 2007).
17. L. Polak and Z. Kotulski*, Sending hidden data through WWW pages: detection and preventions*, pp. 75–89 (2010).
18. Odeh, K. Elleithy, M. Faezipour and E. Abdelfattah, *Novel Steganography over HTML Code. Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering. Lecture Notes in Electrical Engineering* (Springer, Cham, 2015).
19. https://en.wikipedia.org/wiki/Vladilen_F._Minin
20. https://en.wikipedia.org/wiki/Oleg_V._Minin
21. https://en.wikipedia.org/wiki/Igor_V._Minin
22. N. E. Gerasimov, I. V. Minin and O.V. Minin, "Stealthographic protection of intellectual property," in *Www Documents. 8th Int. Scientific Symposium Technomat & Infotel* (Bulgaria, 2007).
23. M. Nighat, Computers in Human Behavior **30**, pp. 648–653 (2014).