ISSN 1512-1232

AN EFFECTIVE METHOD FOR HIDDING DATA IN MICROSOFT WORD DOCUMENTS

Jassam .T. Sarsoh[1], Kadhem .m. Hashem[2], Hayder I. Hendi[3]

[1,2]University of Thi Qar, Education college, Iraq
[3]University of Thi Qar, Computer science and mathematical college, Iraq
Hayderal_khazzai@yahoo.com

ABSTRACT

Most of Microsoft word documents were transfer or shared via internet by using e-mail or any electronic way. This facility is exploited to send any embedded secret data anywhere. In This paper an efficient method was proposed to hide any text massage with in Microsoft word documents.The initial idea of this method is to exploit the addition three spaces on end of paragraph using for hidden data into it by using the color storage bytes.

Two algorithm are proposed and implemented by using PHP language , the first one is to hide the secret message in original text ,and the second is retrieve the original text after the determination of hidden data in it. The comparison of the obtained results between the original text and the text contains the hidden data shows that the two texts are virtually identical.

Key word*: Microsoft word, hidden, Steganography, html*

I. INTRODUCTION

The term steganography is the technique of embedding secret information in a communication channel in such a manner that the very existence of the information is concealed .Steganography techniques have been successfully applied on text files, images, audio and video files [1].in this paper the hidden data was used on text file.

Microsoft word is program correct display of the document on different workstations, even if the two workstations use the same version of Microsoft Word, primarily due to page layout depending on the current printer.A document file format is a text or binary file format for storing documents on a storage media, especially for use by computers. There currently exist a multitude of incompatible document file formats.[2]

2. MICROSOFT OFFICE WORD

Microsoft word is a program which does not guarantee the correct display of the document on different workstations, even if the two workstations use the same version of Microsoft Word, primarily due to page layout depending on the current printer. A document file format is a text or binary file format for storing documents on a storage media ,especially for the use by computer[2]. This means it is possible the document the recipient sees might not be exactly the same as the document the sender sees.

Word document format (.DOC) became the standard of document file formats for Microsoft Office users. Though usually just referred to as "Word Document Format", this term refers primarily to the range of formats used by default in Word version 97-2003. The newer ".docx" extension signifies the Office Open XML international standard for Office documents and is used by Word 2007 as well as by a growing number of applications from other vendors, including OpenOffice.org Writer, an open source word processing program.[3]

ISSN 1512-1232

2.1 CONVERTING MICROSOFT WORD FILES TO HTML

It is necessary to accentuate that Microsoft Word is a versatile program. The problem is that the software may not be the best tool for converting your Word documents to HTML.

It may include different HTML File Types convenient way to get your Word document online is to use the Save as type: Web Page. Then saved HTML file to a web server. There are two issues you should review with this file type.

First issue is That Web Page format appends the information from the File Properties dialog and other descriptive information to the top of the document. These data elements include author, last author, company, document stats and so on. The Web File version is probably fine for company intranets, where users aren't as concerned about privacy. Some of this information could be seen if you emailed the Word file to a co-worker. In contrast, I wouldn't use this format to post your resume on the web especially if you wrote it using a company PC.

The second issue is this HTML format adds tags to the file. One function of these tags is to convert your Microsoft Word style information from your document template. This info also helps you to maintain Word functionality should you need to edit the document. This extra code increases the size of your web page. This may not sound like an issue, but it can be based on your document size. According to sources such as WebSiteOptimization.com, slow response times are one of the most common reasons people leave a site. One part of response time is the web page size.[4]

Microsoft's Word Filtered Web Page Microsoft has another HTML file format called the Web Page, Filtered. This file type strips most of the document information. It also cuts the amount of style codes. Although smaller, this file format still contains numerous references. As mentioned above, some of this coding allows you to edit your work in Word. This file format is best used for final document versions.

3. PHP PROGRAMMING LANGUAGE

Php is the web development language, it was originally name personal homepage (php).in 1994 the software engineer ramous lerdof ,created this language .php is a server side scripting language which can be embedded in html or used as a standalone binary .much of php's syntax is borrowed from c, java and perl with a couple of unique php specific features throw in the goal of this language is to allow web developer to write dynamically generated page quickly .when someone visits the php web page ,of another person ,the web server of that person process the php code. Then sees which parts it needs to show to visitors (content and pictures) and hide the other stuff (file operations, math calculation est.) then translates the php code into html. After the translation into html, it sends the web page to the visitor's web browser .the php can be used to add common header and footers to all the pages of the site or to store from –submitted data in the database .Most of what php does invisible to the end user. if someone looks at a php page ,he will not necessarily be able to tell that it was written in html .[5]

3. PSEUDO-RANDOM NUMBER GENERATOR

A random number generator creates a sequence of randomly distributed numbers. A Pseudo-Random Number Generator creates random numbers as well, but it will create the same sequence of numbers repeatedly. Many algorithms have been developed in an attempt to produce truly random sequences of numbers, with the goal of making it theoretically impossible to predict the next number in the sequence, based on the numbers up to a given point. Unfortunately, the very existence of an algorithm that calculates this number means that the next digit can be predicted **[6]**

4. RELATED WORKS

From a literature point view, some papers were published in the field of the steganography or hiding text.

ISSN 1512-1232

in 2009 Amol .R .Madane and Rashmi Khare[7],took an audio file (wave format)and the converted it to bitmap image using 1D to 2D conversion .the secret message input by the user is hidden at the user specified location

In 2010 ,Nasser hamad [8],used an English text to be hidden into digital grey-scale image .the propose of his research is to embed a maximum text data size into the most suitable image selected among several images based on the binary entropy function can be considered as a powerful tool to select a proper image for a predetermined text.

In February 2011 ,joyshee nath ,sankan das [9] proposed a very effective method to hide some information in some executable file to make the entire process secured ,he had introduced the password when he hide message and while encrypting the secret message ,he had to put some text-key .While hiding secret message in cover file he embeds one byte information in two consecutive bytes of the cover file.

In June 2011, shafik and sankar [1],proposed a novel method for embedding a message in cover audio for secure communication .in his process ,two cover audio files are taken and difference of the amplitude values was calculated. The experimental results show that his technique achieves imperceptible embedding large payload and accurate data retrieval.

Our approach is hide a secret message (text) in Microsoft word document by exploiting the spaces of end paragraph.

5. THE PROPOSED METHOD

Given Microsoft word document and a secret message ,the proposed method aims to hide the secret message with in the Microsoft word by strong each four successive character and the secret message in one space and change the color of the bytes in which these and encryption message in hex are stored. then this process will continue all secret message is hidden in the spaces of end paragraph . this proposed method consists two algorithms, one is to hide the secret message and the other is retrieve the original document.

The following identefiers will be used in the two algorthims

- Text-doc is the given Microsoft word document
- Text –html is the conversion of text –doc to text –html
- Spc is number of spaces in end of paragraphs
- Message is the given secret message
- N is the number of characters of the secret message
- Loc is location of the space in end of paragraph

The following php code statement is used to determine the location of space in the end of paragraph in text –html and change the color of these in order to hide the secret message.

<span style:'color =:#loc'>

To retrieve the secret message text, it is necessary to determent the location In which any space was hide in it. will search the end of paragraph tag "</p> and used # is beginning of hidden message for five space.

5.1 THE HIDING ALGORITHM

This algorithm is to hide the secret message in the given document .it consists the following steps:-

1. Start
2. Input text –doc
3. Convert the number of spc in text-html
4. Calculate the number of spc in text –html

ISSN 1512-1232

5. Input message and encryption it by used the pseudo random generator
6. Calculated the number of byte N in the given message.
7. While (3*spc<N)
   - Choose another text-doc
   - Convert it to text-html
   - Determine its spc

   End while
8. I=0
9. While(i<N and loc of   is found)
   - Store three successive bytes of encryption massage in the loc concerned that
   - I=i+3
10. Convert the obtained text to doc format
11. Output the content of this text-doc
12. End

5.2 THE RETRIEVAL ALGORITHM

This algorithm aims to retrieve the original document. it consists the following steps:

1. Start
2. Read the text-doc in which the secret message is hidden.
3. Convert text-doc to text –html
4. While(not EOF (text-html and loc of   is found)
   - Retrieve the three bytes in loc for five spaces in end of paragraph

   End while
5. Decryption massage by used pseudo random generator
6. Output the plaintext message
7. End

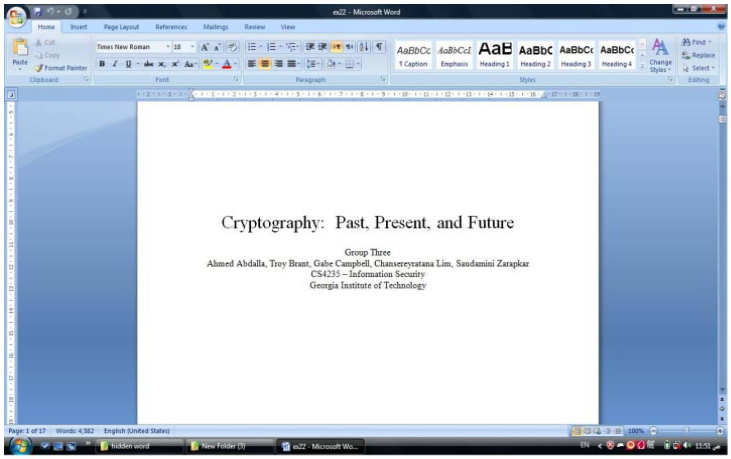
6. EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed method is programmed by using php programming language .this section contains two experiments.

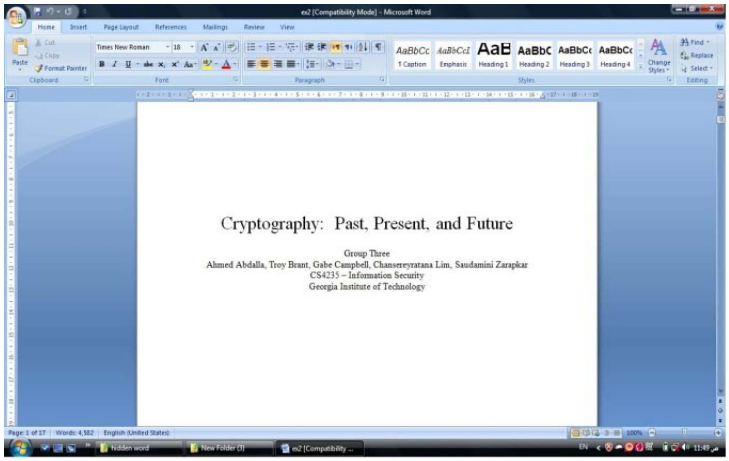
6-1 FIRST EXPERIMENT

In this experiment, chose an English text from the web pages. this text is used as the input Microsoft word document for the hiding algorithm ,and chose a secret message in order to hide it in this given English text ,after that pass the obtained result to the retrieval algorithm .fig(1) shows the original text and the modified text after using our proposed method

ISSN 1512-1232



(a) Original text

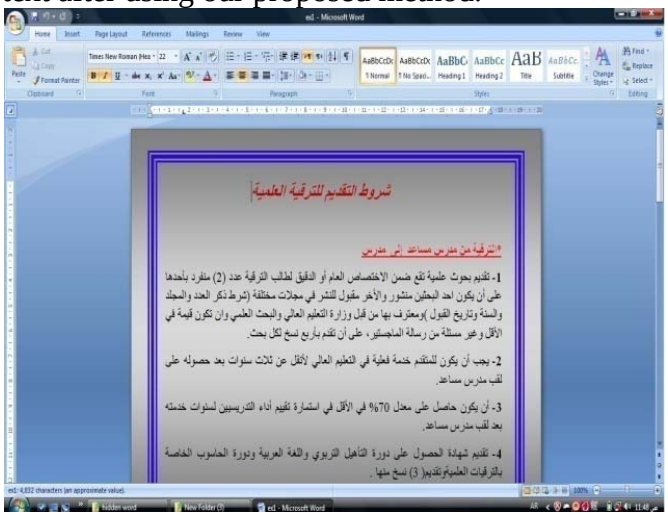


(b) modified text

Fig(1). Show the first experiment

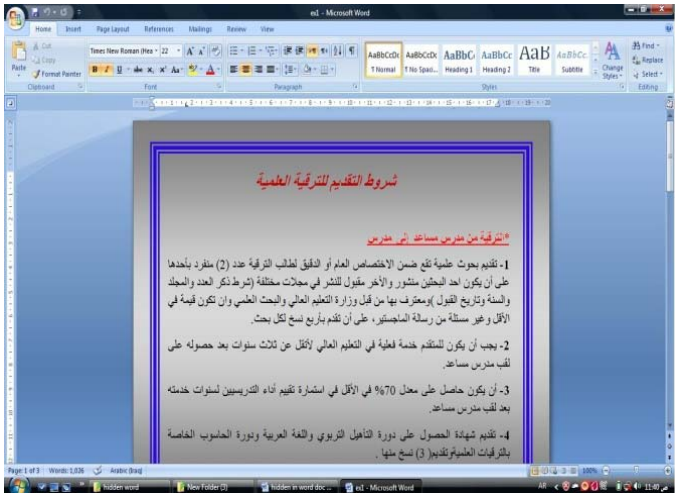
6.3 SECOND EXPERIMENT

In this experiment, an Arabic text in chosen. this text is used as the input Microsoft word document for the hiding algorithm and chose secret message to hide it in the spaces .the obtained text is passed to the retrieval algorithm to retrieve the original Arabic text.fig(2)shows the original and modified text after using our proposed method.



(a) Original text

ISSN 1512-1232



(b) modified text

Fig (2) second experiment

7. CONCLUSION

Using the steganography technique is very useful to hide any secret message in a given cover file .some papers show in related work that this technique can be used to hide a secret message in an image file, audio or video file. our proposed method used to hide any secret text message in a text file .the secret message are hidden in the end paragraph spaces of the Microsoft word document by using the property of the color bytes

Experimental results show that our proposed method give an efficient results .the obtained results shows that there is no different between original document and result document if they were seen virtually by any web page vistor.

REFERENCES

1. Shaff.k , sankar, narayanan , prashanth," A Novel Audio Steganography Scheme using Amplitude Differencing ", IEEE Xplore, ISBN: 978-1-4244-9008-0 vol.10, pp:163-167,2011.
2. Allen, Roy, "A History of the Personal Computer: The People and the Technology", ISBN 978-0-9689108-0-1 , pp.25-26, 2001
3. DeMarco, Jim , "External data is accessed through a connection file, such as an Office Data Connection (DOC) file (.doc), ISBN 978-1-59059-957-0 ,pp.361.
4. "Microsoft Word (.doc) Binary File Format ", 2012-01-22, 2012 Microsoft Corporation
5. Time converse,toyce pack and clarck morgan "php 5 mysql bible",wiley publishing inc,2004
6. Hayder hendi ,"adopted new technique cryptography audio by using genetic algorithm", international Arab conference on e-technology ,pp.7-12 ,2012
7. Amol R.mandane,rashini khare,"time domain steganography",proceeding of the international workshop on machine intellegance research,MIR labs 2009
8. Nasser hamad "hiding text information in digital image base on entropy function",the international arab journal of information technology ,vol.7,no.2,2010
9. Joyshree nath ,sankar das,"advanced stegagraphic approach for hiding encrypted secret message I LSB,LSB+1,LSB+2,LSB+3 bits in non stand cover files ",International Journal of Computer Applications, Volume 14– No.7pp.31-33, February 2011

Article received: 2012-07-10