Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 10 (17): 1916-1922, 2010
ISSN 1812-5654

Investigate the Capability of Applying Hidden Data in Text File: An Overview

[1]A.A. Zaidan, [1]B.B. Zaidan, [2]Ali K. Al-Frajat and [2]Hamid A. Jalab
[1]Faculty of Engineering, Multimedia University, Jalan Multimedia, 63100 Cyberjaya, Malaysia
[2]Faculty of Computer Science and Information Technology,
University Malaya, 50603 Kuala Lumpur, Malaysia

Abstract: The aim of this study is to investigate the methods of steganography using the text file as a cover carrier. The hiding data or Steganography is the art of protecting the information by embedding data in medium carrier, for instants this study illustrates historically this art. The steganography technique has been used mainly to hide secret data within multimedia files and one of used files to hide secret data is the text files. In this study, we have proposed the steganography methods using the text files as a review. The study restricts the weak points of this art by hiding information in text file. As a result of this study, the text based steganography has been discussed and the advantages and disadvantages of using the text file as a cover carrier for steganography has been proposed.

Key words: Data hidden, steganography, text file, hiding in words, hiding in space, hiding in text

INTRODUCTION

All researchers agree that the term is derived from the Greeks and that there were different points of view on some of the words that came out of this term and include some of them below in relation to the sources. This term derived from Greek (Jalab *et al*., 2009; Zaidan *et al*., 2009), means writing a term covered (Covered Writing), or conceal writing (Hiding Writing).

The terminology (Steganography) came from the Greeks and consists of movie (Steganos) means covered or closed and (Graphy) means writing or painting (Naji *at el*., 2009). These mean writing a term covered (Covered Writing). Steganography word can be defined as writing covered (Covered Writing), which is derived from the Greek word (Johnson and Jajodia, 1998).

Thus the definition (Steganography) the art of concealment and transfer data through the data again host or Carrier, but harmful harmless transmitters for those data do not allow any enemy or observers to discover that there is confidential data (Ahmed *et al*., 2010; Majeed *at el*., 2009).

Compared with the information hiding methods intended for images and sounds, there are few methods of hiding information into text. Unfortunately, there is almost no study covers all the methods for hiding information in test such as formatted text such as HTML or XML.

HIDING DATA

One of the latest techniques that have been used in this area by researchers at the Mount Sinai School MOUNT SINAI Medical in New York New York in 1999, as they managed to hide the secret texts in Chromosome Strand human DNA by using a technique called genetic system coverage (Genomic Steganography) and by placing signs resolution to be agreed upon in the nuclei chromosomes and then integrate these with millions sentences and sent to the other end. To extract the secret message is soaking get special distinction sentences used on the other and then placed under the microscope to extract the required text (Clelland *et al*., 1999).

The oldest suthentications on steganography taken from the legendary stories Greeks Herodotus and then back to the fifth century BC, these sources indicate that they felt they fly head of the Messenger and then write the secret letter in the head, leaving hair to grow then be sent to the required which is a re-extraction letter (Johnson and Jajodia, 1998).

Authentications and other writing secret messages on the wood panels and then covered wax and will be hid those writing panels appear free of anything. And they were killing their animals as rabbit example corner confidential letter inside it.

Other means that the common use since the first century AD, invisible inks Invisible Inks, which was able to write a confidential letter with any other non-value-confidential and usually write between lines, for example those rabbis some fruit juices Fruit Juices, milk, urine, vinegar and all these species become dark and visible when exposed to heat the written document.

Then these kinds of inks evolved with the evolution of science chemical was used vehicles carrying chemical characteristics of the same old species with a more

Corresponding Author: A.A. Zaidan, Faculty of Engineering, Multimedia University, Jalan Multimedia, 63100 Cyberjaya, Malaysia

accurate and efficient have been used during the First and Second World Wars in the military secrecy of correspondence. Other technical been used during World War II is sending a message hidden within another message is not relevant and based on the idea of a nomination letters every word of the letter counterfeit representation of characters from the characters letter requested confidentiality (Johnson and Jajodia, 1998).

The earlier application of text based steganography founded during World War II; the Germans would hide data as microdots. This involved photographing the message to be hidden and reducing the size so that it could be used as a period within another document. FBI director J. Edgar Hoover described the use of microdots as the enemy's masterpiece of espionage.

A message sent by a German spy during World War II read: Apparently neutral's protest is thoroughly discounted and ignored. Isman hard hit. Blockade issue affects for pretext embargo on by-products, ejecting suets and vegetable oils. By taking the second letter of every word the hidden message Pershing sails for NY June 1 can be retrieved.

More recent cases of steganography include using special inks to write hidden messages on bank notes and also the entertainment industry using digital watermarking and fingerprinting of audio and video for copyright protection.

Moreover, there are numerous ideas for the same method is used to be more than characters, or take certain words or phrases within the text fake and leaving the rest. Finally, it should be noted that the senior researcher in the area of concealment and science-based organization itself is German Johannes Trithemius between 1462-1526 and the oldest books in the area of coverage Posted by Gaspari Schotti in 1665 in the name of (Steganographyica) and (400) contains a page where all the ideas included (Trithemius).

In this part we will over view some of the related work on the text steganography as the presents in different methods.

In The La Steganography method uses the special form of La word for hiding the data. This word is created by connecting Lam and Alef characters. For hiding bit 0, we use the normal form of word La (ال) by inserting Arabic extension character between Lam and Alef characters. But for hiding bit 1, we use the special form of word La (ﻟا) which has a unique code in the Unicode Standard (its code is FEFB in Unicode hex notation). This method is not limited to electronic documents (e-documents) and can also be used on printed documents.

The method "Dot Steganography", data is hidden in Arabic and Persian texts by using a special characteristic of these languages.

Considering the existence of too many dots in Persian and Arabic characters, in this approach by vertical displacement of the dots, we hide information in the texts. This method does not attract attention and can hide a large volume of information in text.

The other method is using the pointed letters with extension (Kashida in Arabic) to hold secret bit one and the un-pointed letters with extension to hold secret bit zero. Note that letter extension does not have any effect to the writing content. It has a standard character hexadecimal code of 0640 in the Unicode system.

The extension is added before (or after) the pointed letters which can be extended with extension character to hide bit 1 and added before (or after) the un-pointed letters to hide bit 0.

Gutub and Fattani (2007) and Aabed *et al.* (2007) proposed an Arabic text steganography method has been proposed. Gutub and Fattani (2007) proposed steganography approach suitable for Arabic texts. It can be classified under steganography feature coding methods. The approach hides secret information bits within the letters benefiting from their inherited points. To note the specific letters holding secret bits, the scheme considers the two features, the existence of the points in the letters and the redundant Arabic extension character. The author uses the pointed letters with extension to hold the secret bit one and the un-pointed letters with extension to hold zero. This steganography technique is founded attractive to other languages having similar texts to Arabic such as Persian and Urdu. While Aabed *et al.* (2007) embed secret information into text cover media in order to search for new possibilities employing languages other than English. This paper utilizes the advantages of diacritics in Arabic to implement text steganography. Diacritics-or Harakat-in Arabic are used to represent vowel sounds and can be found in many formal and religious documents. The proposed approach uses eight different diacritical symbols in Arabic to hide binary bits in the original cover media. The embedded data are then extracted by reading the diacritics from the document and translating them back to binary.

On the other hand, the whitespace steganography has been used to conceal messages in ASCII text by appending whitespace to the end of lines. Because spaces and tabs are generally not visible in text viewers, the message is effectively hidden from casual observers. And if the built-in encryption is used, the message cannot be read even if it is detected (Takizawa *et al.*, 2001). While Hiroshi *et al.* (2001) proposed an information hiding

method for text which uses text as not paper images but sequences of characters. That means that linguistic expressions are altered to hide hidden information while the meaning of the text is preserved. Our target text is written in Japanese in the field of software manual and document for user agreement of the software. Actually, this method hides information into text by paraphrasing with a dictionary which consists of pairs of expressions having the same meaning.

Some methods for hiding data into text process texts as image essentially. Those methods have a characteristic that the copy of printed matter has the same secret data as original. There are some methods hiding data into text data using character codes. For example, there is the method appending whitespace to the end of lines (Inoue *at el*., 2001) and the method changing the start position of new lines (Inoue *at el*., 2001).

Another approach is to handle texts not as paper image but sequence of characters (Hiroshi *at el*., 2001). The method intends that linguistic expressions are altered to hide secrets while the meaning of the text preserved.
For Structured documents, there is a study on PDF and PostScript (Shibuya *et al*., 1998). The method, changes to embed secret data are done not to change the final output.

Al-Azawi and Fadhil (2010) hides information by inserting extension characters (Kashida) at suitable word positions. We insert extension character in a word position to hold secret bit one and leaving position empty to hold secret bit zero. The Huffman compression algorithm is used to sconvert the embedding message into a compressed binary fonn and an Arabic text steganography technique based on character extensions is used to insert the compressed binary into the determined positions of the words in the cover text.

HIDDEN DATA IN TEXT FILE

Hidden in text: The most common methods of concealment and simplest is Switches (Binary Digit) known briefly as (bit), least significant known as (LSB), Where it is altered binary digit characters to the message characters to be hidden, after conversion of such characters to byte as well as the American Standard Cod for Information Interchange (ASCII).

We cannot use this method here because switching binary digit might lead to an increase or decrease the value of letter by (1); this leads to the advance of this letter with the neighbor letter, for example the letter (C) in English represent in binary (100 0011) with replacement the Least Significant bit the binary value become (100 0010) which is represent (B) in English, that will make the carrier text become a meaningless, which Denies the goal of hidden technique, Therefore resort to other means which exploit spaces between strings and words in the carrier text (Zaidan and Zaidan, 2009).

LINE SHIFT CODING PROTOCOL

In line shift coding, we simply shift various lines inside the document up or down by a small fraction (such as 1/300th of an inch) according to the codebook. The shifted lines are undetectable by humans because it is only a small fraction but is detectable when the computer measures the distances between each of the lines. Differential encoding techniques are normally used in this protocol, meaning if you shift a line the adjacent lines are not moved. These lines will become a control so that the computer can measure the distances between them.

By finding out whether a line has been shifted up or down we can represent a single bit, 0 or 1. And if we put the whole document together, we can embed a number of bits and therefore have the ability to hide large information.

WORD SHIFT CODING PROTOCOL

The word shift coding protocol is based on the same principle as the line shift coding protocol. The main difference is instead of shifting lines up or down, we shift words left or right. This is also known as the justification of the document. The codebook will simply tell the encoder which of the words is to be shifted and whether it is a left or a right shift. Again, the decoding technique is measuring the spaces between each word and a left shift could represent a 0 bit and a right bit representing a 1 bit.

- The quick brown fox jumps over the lazy dog
- The quick brown fox jumps over the lazy dog

In this example the first line uses normal spacing while the second has had each word shifted left or right by 0.5 points in order to encode the sequence 01000001, that is 65, the ASCII character code for A. Without having the original for comparison it is likely that this may not be noticed and the shifting could be even smaller to make it less noticeable.

FEATURE CODING PROTOCOL

In feature coding, there is a slight difference with the above protocols and this is that the document is passed through a parser where it examines the document and it

automatically builds a codebook specific to that document. It will pick out all the features that it thinks it can use to hide information and each of these will be marked into the document. This can use a number of different characteristics such as the height of certain characters, the dots above i and j and the horizontal line length of letters such as f and t. Line shifting and word shifting techniques can also be used to increase the amount of data that can be hidden.

TEXT CONTENT

Another way of hiding information is to conceal it in what seems to be inconspicuous text. The grammar within the text can be used to store information. It is possible to change sentences to store information and keep the original meaning. Text Hide is a program, which incorporates this technique to hide secret messages. A simple example is:

The auto drives fast on a slippery road over the hill. Changed to:

Over the slope the car travels quickly on an ice-covered street.

Another way of using text itself is to use random words as a means of encoding information. Different words can be given different values. Of course this would be easy to spot but there are clever implementations, such as SpamMimic which creates a spam email that contains a secret message. As spam usually has poor grammar, it is far easier for it to escape notice. The following extract from a spam email encodes the phrase "I'm having great time learning about computer security.

Dear Friend, Especially for you-this red-hot intelligence. We will comply with all removal requests. This mail is being sent in compliance with Senate bill 2116, Title 9; Section 303 ! THIS IS NOT A GET RICH SCHEME. Why work for somebody else when you can become rich inside 57 weeks. Have you ever noticed most everyone has a cell phone and people love convenience. Well, now is your chance to capitalize on this. WE will help YOU SELL MORE and sell more! You are guaranteed to succeed because we take all the risk! But don't believe us. Ms Simpson of Washington tried us and says "My only problem now is where to park all my cars. This offer is 100% legal. You will blame yourself forever if you don't order now! Sign up a friend and you'll get a discount of 50%. Thank-you for your serious consideration of our offer. Dear Decision maker;

Thank-you for your interest in our briefing. If you are not interested in our publications and wish to be removed from our lists, simply do NOT respond and ignore this mail! This mail is being sent in compliance with Senate bill 1623; Title 6; Section 304 ! THIS

IS NOT A GET RICH SCHEME! Why work for somebody else when you can …

A very basic form of steganography makes use of a cipher. A cipher is basically a key which can be used to decode some data to retrieve a secret hidden message. Sir Francis Bacon created one in the 16th Century (Shirali-Shahreza and Shirali-Shahreza, 2008) using messages with two different type faces, one bolder than the other. By looking at the positions of the bold characters in relation to the rest of the text, a secret message could be decoded. There are many other different ciphers which could be used to the same effect.

DOTS STEGANOGRAPHY METHOD

This method depends on the points in the Arabic letters. This large number of points in Arabic letters made the points in any given Arabic text can be utilized for steganography information security. The dots letters is used to hide bits. The dots slightly shifted up more than normal to represent the hidden bit 1 and kept the pointed character normal to hide o. In this method, robustness is weak since it depends on using same fixed font, where using different font to produce unknown letters (Bennett, 2004).

ARABIC DIACRITICS METHOD

This method utilizes the advantages of diacritics in Arabic to implement text steganography. Arabic text uses eight different diacritical symbols and this method uses the most frequent diacritical symbol Fatha. One of these methods, at start a fully diacritised Arabic text is used as cover media. To hide a bit 1, all diacritics are removed from the cover media until a Fatha is found and to hide a bit 0, the first non Fatha diacritic is kept. That means each Fatha represents 1 and other diacritic represents O. The overall process is repeated for as long as there are bits remaining to be hidden (Aabed *et al*., 2007). We need to note that diacritics approach, as well as the Kashidah approach, hiding a bit is equivalent to inserting a character (a diacritic mark or a Kashidah) (Gutub and Fattani, 2007). The main advantages of this method are: provides the highest capacity, fast, does not require large computational power and can be implemented manually. While the main disadvantages are: suspicions raise since, it is uncommon nowadays to send diacritized text, the output text has a fixed frame due to the use of only one

font and the information is lost in case of retyping (Aabed *et al*., 2007; Gutub and Fattani, 2007).

ARABIC UNICODE TEXTS USING PSEUDO-SPACE AND PSEUDO CONNECTION

In Arabic unicode texts, there are two characters, pseudo-space (ZWNJ-zero with non joiner) and pseudo-connection (ZWJ-zero with joiner) characters which are, respectively prevents Arabic letters from joining or forces them to join together. This method, first looks if the letter in a word connected to the next letter or not. To hide bit 1, ZWJ letter is inserted between letters if connected or ZWNJ letter is inserted between letters if not connected and do not add anything for hiding bit O. This method is not dependent on any special format and can hide information in numerous formats such as HTML pages, Microsoft Word Documents and also capable to hide a bit of information in each letter (Shirali-Shahreza and Shirali-Shahreza, 2008).

HIDDEN IN SPACE

In this technical parts of the secret letter tight in the spaces of the carrier letter these technical include ways:

- First method: embed one or two space after each set of phrases for the carrier letter
- The second method: contains the same way of the first method except that the process is embedding after every word of the carrier letter words

The first method more efficient for lack of change compared with the original text although it is not efficient in terms of the amount of data that possible to embed, the way in the second method could embed more quantitative data but more likely for observation and perception compared to the original text (Zaidan and Zaidan, 2009; Johnson and Jajodia, 1998). Example of text steganograph is shown in Fig. 1.

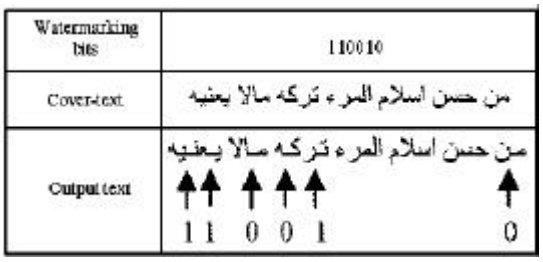

Fig. 1: Example on text steganography proposed by Gutub and Fattani (2007) (Hidding in space)

HIDDEN IN WORDS

This technique was used during the Second World War, sending a message hidden within another message which is not relevant, the role of this idea is nomination letters in every word of the Carrier letter from the characters confidential letter to be sent (Eltahir *at el*., 2009).

One of these examples was the letter that was sent by Deutsch Spy during the Second World War Apparently neutrals protest is thoroughly discounted and ignored. Is man hard hit Blockade issue affects pretext for embargo on by-products, ejecting suet's and vegetable oils.

By analyzing this letter crafts taking the second letter in each word it will show us the following secret letter Perishing sails from NY June1.

This method is very complex and not efficient in terms of application in the computer system because it requires an expert system in the field of language through which to choose the suitable word and build the sentence required, as well as a huge database requires (Zaidan and Zaidan, 2009).

DISADVANTAGE HIDDEN IN TEXT

Methods of concealment in the text weak and inefficient not suitable for the application and the main disadvantages include:

- Need large text to hide small message
- Some methods need to complex techniques for the application like method of hidden information in words
- The possibility of noting changes in the carrier letter compared with the original
- Probability crash system of concealment in the case presented the letter using one service applications such as (word) as a result of it contains some automatic format, which leads to change along the spaces between words (for the English version)

ASCII

American Standard Code for Information Interchange (ASCII) is a coding standard that can be used for interchanging information, if the information is expressed mainly by the written form of English words. It is implemented as a character-encoding scheme based on the ordering of the English alphabet. ASCII codes represent text in computers, communications equipment and other devices that work with text. Most modern

character-encoding schemes which support many more characters than did the original have a historical basis in ASCII.

- Concealment is limited to text message with difficulty concealing other types of messages like equations, charts, images and sounds
- In additional there's a solution; Office 2003/ XP Remove Hidden Data Add-in, It's a free add-in for Office 2003 and XP that clears hidden data from Word, Excel and Power point files forever

FORMATTED DATA WRITTEN IN MARK-UP LANGUAGES

This technique considers that the number of applications will increase which use not only plain-text but formatted data written in mark-up languages such as SGML or HTML. Nowadays, XML is known as the universal format for structured documents and data and used as the basic technology for exchanging information on the Web. The importance of security on XML is growing more.

Although still HTML has been widely used to describe Web pages, XML pages can be created and browsed by using major browser.

Information hiding methods using XML satisfies the following conditions.

- Cover data is either or the entire XML document, DTD, XSL and CSS
- he cover data is changed to embed data while the meaning is preserved in stego data.

Inoue *et al*. (2001) proposed hiding information in white spaces in tags. Representation of a tag is either including some white spaces before close blacket, or no white space. The author embedded the data preserving all meaning of original document by inserting or deleting spaces (Inoue *et al*., 2001).

Another way presented by Inoue *et al*. (2001) where the embedding of the secret data in XML documents by exchanging of the appearing order of elements. One bit of data can be hidden in the documents per an exchange of two elements (Inoue *et al*., 2001).

MAKING FONT COLOUR ALMOST MATCH BACKGROUND

After thousands of webmasters learned the hard way that Google was able to detect and penalize sites which

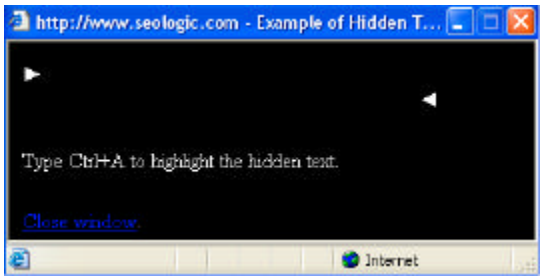

Fig. 2: The site before highlighting the page

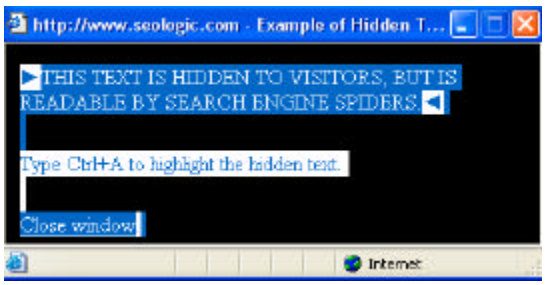

Fig. 3: The site after highlighting the page

used Methods 1 and 2 above, many altered their methods slightly. Instead of setting the font colour to match the background colour exactly, they set their font colours to almost match the background colour (Fig. 2, 3).

It is easy for a human to review your page and find your hidden text by simply selecting and highlighting the whole page. To see the hidden text we've just created above, click on Edit at the top of your browser window and select Select All. Or, just hit Ctrl+A to do it quickly to any page. When you select all, invisible text will appear which was previously hidden because the page will change colour slightly-it will be highlighted. Alternatively, you can use your mouse to select the invisible text between the arrows and it will appear highlighted.

The idea behind this Method is that the webmaster or SEO consultant will be able to argue that they aren't specifically violating the search engines' terms of service. After all, the text isn't quite invisible. Some also believe that they are thwarting the search engines' software detection systems by changing the colour of the text slightly. Neither of these assumptions is correct.

CONCLUSION

In this study, we have clarified the intended knowledge of data hidden. In addition, we have presented the history of the Steganography since ancient times until the present day. One of the challenges in this article

reviewing the most important methods in the text file that used in this field. Finally, we demonstrate the disadvantage points of the hiding in text.

ACKNOWLEDGMENT

This research has been funded by the University of Malaya, under the grant number (P0033/2010A). The author would like to take this opportunity to thank and acknowledge all his friends and associates who had offered him the much needed assistance and encouragement from the start to the end of the research period.

REFERENCES

Aabed, M.A., S.M. Awaideh, A.M. Elshafei and A.A. Gutub, 2007. Arabic diacritics based steganography. Proceedings of the International Conference on Signal Processing and Communications, Nov. 24-27, Dubai, UAE, pp: 756-759.

Ahmed, M.A., M.L.M. Kiah, B.B. Zaidan and A.A. Zaidan, 2010. A novel embedding method to increase capacity and robustness of low-bit encoding audio steganography technique using noise gate software logic algorithm. J. Applied Sci., 10: 59-64.

Al-Azawi, A.F. and M.A. Fadhil, 2010. Arabic text steganography using kashida extensions with huffman code. J. Applied Sci., 10: 436-439.

Bennett, K., 2004. Linguistic steganography: Survey, analysis and robustness concerns for hiding information in text. CERIAS Technical Report, Purdue University, West Lafayette, IN 47907-2086. https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/2004-13.pdf.

Clelland, C.T., V. Risca and C. Bancroft, 1999. Hiding messages in DNA microdots. Nature, 399: 533-534.

Eltahir, M.E., M.L.M. Kiah, B.B. Zaidan and A.A. Zaidan, 2009. High rate video streaming steganography. Proceedings of the 2009 International Conference on Future Computer and Communication, April 03-05, IEEE Computer Society, Kuala Lumpur, Malaysia, pp: 550-553.

Gutub, A.A.A. and M.M. Fattani, 2007. A novel Arabic text steganography method using letter points and extensions. World Acad. Sci. Eng. Technol., 27: 28-31.

Hiroshi, N., S. Koji, M. Tsutomu, K. Takeshi, K. Shuji, M. Kyoko and M. Ichiro, 2001. Meaning preserving information hiding. Japanese text case. Trans. Inform. Process. Soc. Jap., 42: 2339-2350.

Inoue, S., K. Makino, I. Murase, O. Takizawa and T. Matsumoto, 2001. A proposal on information hiding methods using XML. Proceedings of the 1st Workshop on NLP and XML, Nov. 2001, Japan. http://takizawa.ne.jp/nlp_xml.pdf.

Jalab, H., A. Zaidan and B.B. Zaidan, 2009. Frame selected approach for hiding data within MPEG video using bit plane complexity segmentation. J. Comput., 1: 108-113.

Johnson, N.F. and S. Jajodia, 1998. Steganalysis: The investigation of hidden information. Proceedings of IEEE Information Technology Conference, Sept. 1-3, New York, USA., pp: 113-116.

Majeed, A., M.L.M. Kiah, H.T. Madhloom, B.B. Zaidan and A.A. Zaidan, 2009. Novel approach for high secure and high rate data hidden in the image using image texture analysis. Int. J. Eng. Technol., 1: 63-69.

Naji, A.W., A.A. Zaidan and B.B. Zaidan, 2009. Challenges of hidden data in the unused area two within executable files. J. Comput. Sci., 5: 890-897.

Shibuya, R., Y. Kaji and T. Kasami, 1998. Digital watermarking for PostScript and PDF documents. Symposium on Cryptography and Information Security, SCIS98-9.2.E, Jan.

Shirali-Shahreza, M. and S. Shirali-Shahreza, 2008. High capacity persian/arabic text steganography. J. Applied Sci., 8: 4173-4179.

Takizawa, O., A. Yamamura, H. Nakagawa, 2001. A proposal of steganography on plain text and XML. Proceedings of the 7th Annual Meeting of the Association for Natural Language Processing, March 2001, Tokyo, pp: 135-138.

Zaidan, A. and B. Zaidan, 2009. Novel approach for high secure data hidden in MPEG video using public key infrastructure. Int. J. Comput. Network Security, 1: 1985-1993.

Zaidan, A., B. Zaidan and F. Othman, 2009. New technique of hidden data in pe-file with in unused area one. Int. J. Comput. Electrical Eng., 1: 1793-8198.