Computers in Human Behavior 30 (2014) 648–653

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Copyright for web content using invisible text watermarking

Nighat Mir *

Computer Science Department, College of Engineering, Effat University, P.O. Box 6290, Jeddah 21442, Saudi Arabia

ARTICLE INFO

Article history:
Available online 5 September 2013

Keywords:
Digital watermarking
HTML
Copyrights
HASH
Invisible
Embedding

ABSTRACT

Digital watermarking is a copyright protection technique used to embed specific data in a cover file to prevent illegal use. In this research invisible digital watermarking based on the text information contained in a webpage has been proposed. Watermarks are based on predefined semantic and syntactic rules, which are encrypted and then converted into whitespace using binary controlled characters before embedding into a webpage. Structural means of HTML (Hyper Text Markup Language) are used as a cover file to embed the formulated watermarks. Proposed system has been validated against various attacks to find optimum robustness.

1. Introduction and background

Information and security are two codependent disciplines and with the advent growth in internet and related technologies have brought many challenges to manage security. In this global tech era information has shaped into electronic-information which heavily floats on network every instant and everywhere. Internet influx has provided flexible means of sharing online information in an economical way and attracted many writers over this short period of time and in the meantime has also brought a huge demand of intellectual copyrights protection mechanisms. Electronic information sharing brings many threats to its integrity and authenticity. Security of information is to assure its protection against various intimidations like reusing, disrupting, redistributing, tampering, authenticating, copying, modifying and forgery of information in an illicit way for dishonest purposes.

Different methods have been studied for multimedia objects but few are available for securing textual information without altering its integrity. Web based attacks have been a very common practice in recent years and hence need strong security mechanisms for the sake of secret communication. Beside other data types available on web, text is the most dominant part and is of utmost importance of securing it. Therefore, it is urgently required to have rights protection solutions which remain attached to the text even if it is re-produced, edited, and modified.

Syntactic and semantic methods based on the language constructs have been used for text watermarking and hence can be extended towards the web content as text is the most leading part of a web page. Cover file (an object in which a secret information can be carried or embedded) to a large bandwidth which is capable of developing and embedding secret digital codes for copyrights protection. Web pages have a privilege of comparatively having greater amount of bandwidth which can be utilized in best possible ways for securing the digital codes.

Most of the methodologies studied for digital text watermarking depend on the content itself and do not address the medium. Web watermarking addresses the issue of protecting not only the content but the carrying medium itself by taking the construction elements of the page into mind. A web page can be developed or designed in many different tools but eventually it is translated into HTML (Hyper Text Markup Language) code by any associated browser which is the basic construction mechanism. Whereas XML (eXtensible Markup Language) is used massively for trading information on internet. These basic construction tools can be securely integrated with the web content for copyright protection.

Growths of internet use and copy culture present a challenge for copyrights of original information on web. Internet is a centralized global location intended to exchange information rather than copying from it. However, with the increasing threats to online information it is imperative to extend a research into different security techniques. It is quoted in Benjamin (1969) about photography that "one can make any number of prints from a negative and to ask for the authentic print which makes no sense". This generalization can be applied nowadays to the internet where mostly people do not ask or care about the original or copy. Moreover, one of the reasons is the diverse culture of cyberspace inhabitant who never knew the internet world before knowing new internetusers. Therefore, any online information is considered as free and users instead of exchanging are just taking as information. Literature indicates that internet is raising a generation where students think

* Tel.: +966 530149748.
E-mail addresses: nighatmir@gmail.com, nmir@effatuniversity.edu.sa

0747-5632/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.chb.2013.07.040

that any piece of online information can be used readily (Renard, 1999).

Copyright protection system preserves the rights of an author for life time and 50 years after his death. It allows authors to reproduce, enhance, sell or loan someone their protected work (Gerasimov, Minin, & Minin, 2007).

Digitalization of information on internet allows people an easy access to the information to copy and to transmit data using different mediums (Arnold, Schmucker, & Wolthusen, 2003). As such, it requires ensuring enforcement and protection of electronic data copyrights (Katzenbeisser & Petitcolas, 2000).

Besides protecting the information, methodologies to protect the carrier of this information also needs to be secured. The same has been highlighted in this research by offering sanctuary of web pages based on their content. To protect copyright, digital watermarking has been used and to secure watermarks, cryptography has been combined further for the embedding and extraction process.

Watermarking is a robust method against various attacks. Even if the existence of hidden information is known, it is extremely difficult to know or decode the watermarked information. Digital watermarks should be invisible and maintain the data integrity, robustness, convenient watermark extraction that is able to withstand its existence in a cover file which carries the secret information. Therefore, all these features have been adequately addressed in this research to offer a digital web watermarking.

Basic security parameters authenticity, integrity and confidentiality are cores to be achieved by using digital watermarking (Zhao & Koch, 1995). Security techniques have mostly been applied to video, audio and image based content and less focused on text due to its less redundant structure (Mir & Hussain, 2011). Numerous strategies have been proposed for online information to filter unsafe web pages and several frameworks have been studied for web text against different attacks (Mulwad, Li, Joshi, Finin, & Viswanathan, 2011; Zhu, Xu, Jiang, & Chen, 2010).

Importance of text data has an obvious relation with information security and in the meantime should preserve the textual meaning due to its heterogeneous and greater value set (Martínez, Sanchez, Valls, & Batet, 2012). Research in the field of information security is evolving interest with the development of technology for researchers and academicians (Zhao & Xue, 2009). Whereas objective is to protect the ownership of an author by discouraging various online thefts (Turner, 1990).

Text watermarking, text stenagography and cryptography are known mechanisms for providing security, each in its own and different manner. These have been used both individually and combined in research for developing new security mechanisms. Digital watermarking and cryptography have been clustered to validate webpage fidelity by calculating the web visit counts and digital fingerprinting for the purpose of securing the watermarks (Bo, Wei, Yuan-Yuan, Ying-Zhi, & Dong-Dong, 2009; Zhou & Li, 2009). In this research web watermarking and cryptography have been complimented with each other to offer copyright protection for online text data.

Digital Watermarking has been proposed in literature as a solution to protect authenticates and secures information for various applications (Lu, 2005). Various aspects of text have been addressed using the syntactic (Atallah, Raskin, & Crogan, 2001), semantic (Atallah et al., 2002) and structural (Li & Dong, 2008; Lu et al., 2009) features to provide text watermarking. Web watermarking relies on semantic and structural features of different mark up languages.

Special stealthographic techniques allowed embedding invisible codes with a web page to HTML invisible codes to address the security problem (Gerasimov et al., 2007). White space characters do not have a visual appearance and mainly used to provide readability by adding spaces between sentences, words, character or lines. Many programming languages ignore white spaces during the compilation process of a code. Mostly it is added by the spacebar, tab, enter or new line character to provide vertical or horizontal style in typing.

Behind these simple ways of adding whitespace there are some values or unicode characters assigned to these white spaces in computer memory. These are usually called "invisible" or "no face" characters. Most of the character encoding schemes is based on ASCII (American Standard Code for Information Interchange) which provides a textual representation of text in computers. Invisible control characters are represented by Hex-Code 20 and represented by U+number in unicode e.g. U200A, U2422, U202F and so on (http://www.ietf.org/rfc/rfc2822.txt).

ASCII or Unicode do not describe the procedure for formatting or textual appearance in a document where markup languages such as HTML, XML, and SGML are very much concerned more in the structure, layout and formatting aspects. Hence, white space plays a significant role in the appearance of these markup languages in contrast to unicode schemes.

White spaces are treated differently in HTML from other characters and a sequence of white space is generally collapsed by using these as a single space. Frequently preserving white spaces are not as important for web authors as they can apply indentations by using relevant tags. On the other hand, sometimes preserving whitespaces become very crucial and for this HTML provide a tag such as <pre> that preserves all types of spaces (e.g. spaces, tabs and newlines) (http://www-sul.stanford.edu/tools/tutorials/html2.0/whitespace.html). Also like HTML, XML it also ignores spaces while parsing where a special attribute named xml:space with values "default" or "preserve" has been provided by XML to preserve the spaces in an element (http://www.w3.org/TR/2004/REC-xml11-20040204/#sec-white-space).

Digital Watermarking provides copyright protection where watermark is embedded into text files. Text as a simple and convenient mode of communication involves certain properties to be addressed e.g. text pattern, binary nature, fluency, language rules, etc. and text watermarking preserves the language rules.

Previous techniques which have been proposed are usually categorized as image (Atallah, McDonough, Nirenburg, & Raskin, 2000; Brassil, Low, & Maxemchuk, 1999; Brassil, Low, Maxemchuk, & Gorman, 1995), and syntactic(Topkara, Topkara, & Atallah, 2006) and semantic(Topkara, 2007; Topkara, Topraka, & Atallah, 2007) approach for text watermarking mainly used for text documents are not considering much on web based text. Text has mainly taken in form of images due to the challenge of less redundancy, typing errors, nouns, verbs and tree structures of content.

2. Methodology

Proposed methodology results invisible watermark to be embedded into a webpage using the structural features of HTML (Hyper Text Markup) language. Web textual content is parsed to generate watermarks, based on predefined a semantic and syntactic rule which goes through a cryptographic cycle (using HASH algorithm) to generate a hash function which is replaced using invisible control characters to generate on invisible set of watermark.

In this research, language tests have been conducted using English language features. However, application is flexible enough to take any other language constructs, thus ideas can easily be extended further. In this research a specific list of verbs (is, are), articles (a, an), and frequently occurring initial letters (wh, th) of English language have been considered only. HASH is used for cryptographic purposes and unicode controlled characters for

invisibility of watermark. HTML <meta> tag has been selected as a cover file source location to embed the invisible watermark.

2.1.1. Process

The process consists of four basic steps for generating watermarks, producing corresponding hash functions for watermarks; converting the hash value into 8 digits value; converting into white spaces; then embedding into a webpage. The algorithm shows the programmed process for the proposed system.

2.2.2. Algorithm[1]

1. I/P: webpage text //taking as input web text to scan/read for given rules
2. O/P: whitespace // output is invisible watermarks
3. repeat //to generate watermarks
   {
   a. step 1
   b. TW = compute based on semantic/syntactic rules
   c. cover file: TW*FoOcc
   d. while(!EOF)
   e. EN_TW = apply HASH(coverfile)
   do //to generate hash value
   {
   get TW — > 512 bits chunk (MsgChk)
   % MsgChk — > 16 MsgChk(32 bits)
   initializeH_Value
   Trim H_Value — > 8 digits
   }
   returnH_Value
4: H_Value
   repeat (for i <= 8) //repeat for 8 digits
   {
   start = u200A
   replace = u202F
   end = u205F
   }
   return invisible watermarks(IV_W) //whitespaces (8 digits)
   }
   f. embed IV_W — > (HTML<meta key= IV_W) />
5: end

Four main points of algorithm are explained below in steps:

Step 1. Generating Watermarks
1. Read a URL and scan the text
2. Based on semantic, syntactic or any rules; generate a watermark
3. Generate HASH for the corresponding watermark

Step 2. Generating HASH function for watermark
1. Take the message in form of String values
2. Process this message into 512 bits consecutive chunk (MsgChk)
3. Break this 512 chuck (MsgCHK) into 16 chuncks of 32 bit words
4. For MsgCHK, initialize the HASH value
5. Get the MsgCHK's HASH
6. Trim the MsgCHK into 8 digits
   a. Take the variable length MsgCHK HASH
   b. Convert into binary digits
   c. Trim the values
   d. Get the concatenated fixed length 8-digit HASH value

Step 3. Converting 8-digit HASH into White space using Controlled Characters
1. Take the 8-digit HASH value
2. Read the first digit using " u200A" convert the first digit into white space using " u202F" and end the process by using " u205F" controlled character.
3. Repeat the process 8 times
4. Embed the white space HASHED characters into a web page

Step 4. Embedding the Watermark
1. Take the white space HASHed watermark
2. Embed it into meta tag of HTML source code

3. System design

Figures shown below indicate the details of the proposed system in a pictorial process. Fig. 1 shows the details of digital watermark generation, encryption, conversion into white spaces and the final embedding process. A URL is subjected to the system, which feeds the text data to the application. Predefined rules for a language have been extracted from the webpage textual content that constructs the initial watermarks. These initial watermarks are encrypted using the HASH function which is further trimmed to produce a fixed length watermark of 8 digits numbers string. During the trimming process, information is preserved hence no loss of data occurs. Generated hashed string is first converted into binary codes and similar values are taken once, which are reconstructed through a reverse process to the original value.

Unicode control characters have been used in this research to craft the invisible watermarks. These unicode characters are known as "no face" characters, as these appear in the form of white spaces. The purpose of hiding the watermarks is the fact that the HTML source code is visible to users and watermarks are prone to deletion attacks. The process is made more robust by applying invisible digital watermarking. Three unicodes used in this research are u200A, u202F and u205F.

Hashed 8 digits string is subjected to the process of conversion to white spaces, where the process repeats 8 times for each digit. For each digit mentioned unicode controlled have been used for the conversion process, first character u200A has been used to read the value, second character u202F for actual conversion and third character u205F for the ending and then it reads the next digit following the same process. A whitespace watermark consisting of 8 digits has been produced as a result of invisible watermark to be embedded to a webpage. HTML <meta> tag has been used as a source location, as by definition it is used for specifying any description like author's information, key, date etc.

Fig. 2 shows the extraction and validation process for the watermark. First it takes whitespaces from the <meta> tag of HTML source code and then reads the corresponding values before decrypting the watermarks. A hash code is recovered and match with the original for the validation process. Original and generated watermarks are then ready to be verified for originality based on the level of similarity between the two.

Experiments for this research have shown the exact reconstruction and match to the original.

[1] All variables/notations used in the algorithm have been explained in Appendix section.

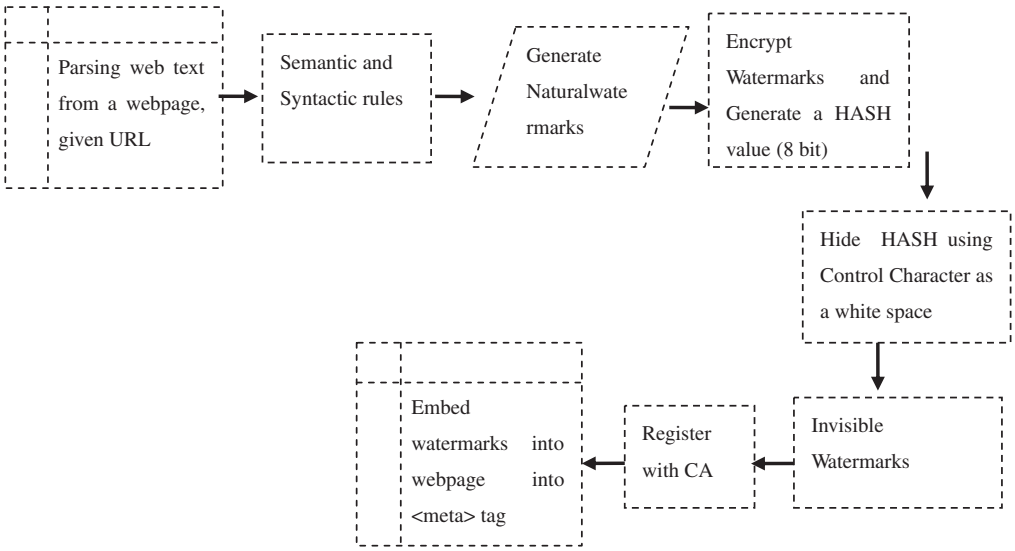

Fig. 1. Watermark generation, embedding and conversion process.

4. Experimental results

Results shown below have been tested on the link (http://en.wikipedia.org/wiki/Cryptographic_hash_function) and features for generating watermarks checked for this website are (this, that, the, which, is, are, a, of). Fig. 3 shows the steps performed for the application to verify the proposed approach. Step 1 and 2 shows the process of taking URL of a web page to read the text. Step 3 gives rules to be used for generating watermarks from a web page. Step 4 shows the count of scanned watermarks from a web page according to the given rules in step 3. Steps 5 and 6 give a view of generating hash value which is saved as a text file and shown in step 7. Generated hash value is given to the application to generate a white space character, as shown in the step 7 and the highlighted gray color part is the no face character in form of while space, which is finally embedded to the meta tag of a webpage as white space 8 bits character.

Few test websites demonstrated in Table 1 to show the results of the implemented solution. Three web pages from Wikipedia shown with the number of occurrences for each verb and article defined for the research along with their hash value, whereas whitespaces have been highlighted in grey color for visibility in the last column.

5. Conclusion and discussion

A novel robust web based text watermarking algorithm is proposed in this paper. To secure text, a webpage invisible watermarking has been implemented by combining a cryptography algorithm. Special no face control characters are used to offer the invisibility to convert the watermarks into whitespaces which are not visible and robust against basic modification and deletion attacks. White spaces are usually ignored by many programming languages and so on markup languages where using the defined rules of markup languages these whitespaces can be preserved at their actual positions. A defined descriptive tag is used in this research to embed the invisible watermark.

In this research specific rules of English language have been used. However, proposed idea is independent of language and can be easily extended for any other language by exploiting the same or other semantic, syntactic and structural features. Features mainly could be based on the high frequency of occurrence. Tested language can also be extended for similar or any other language aspects. HTML has been used in this research to embed the watermark and this can be extended for other markup languages like XML and SGML; as XML plays a major role in trading information over internet. A basic descriptive HTML tag has been used in this research and many other structural aspects can be considered for embedding the

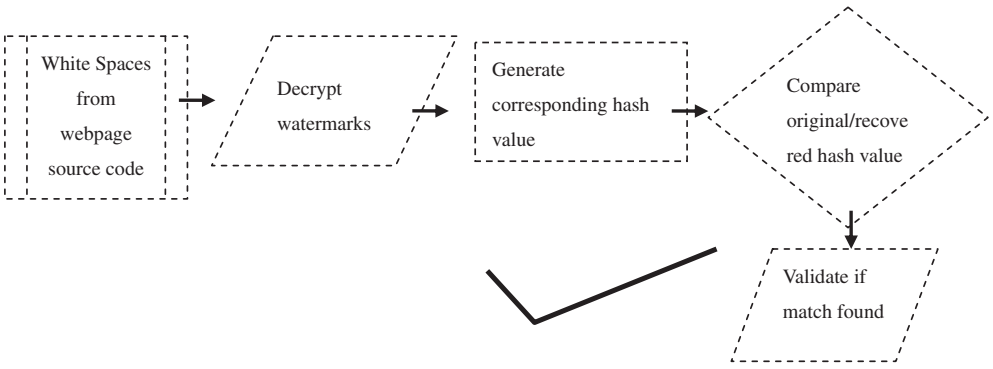

Fig. 2. Extraction and validation process.

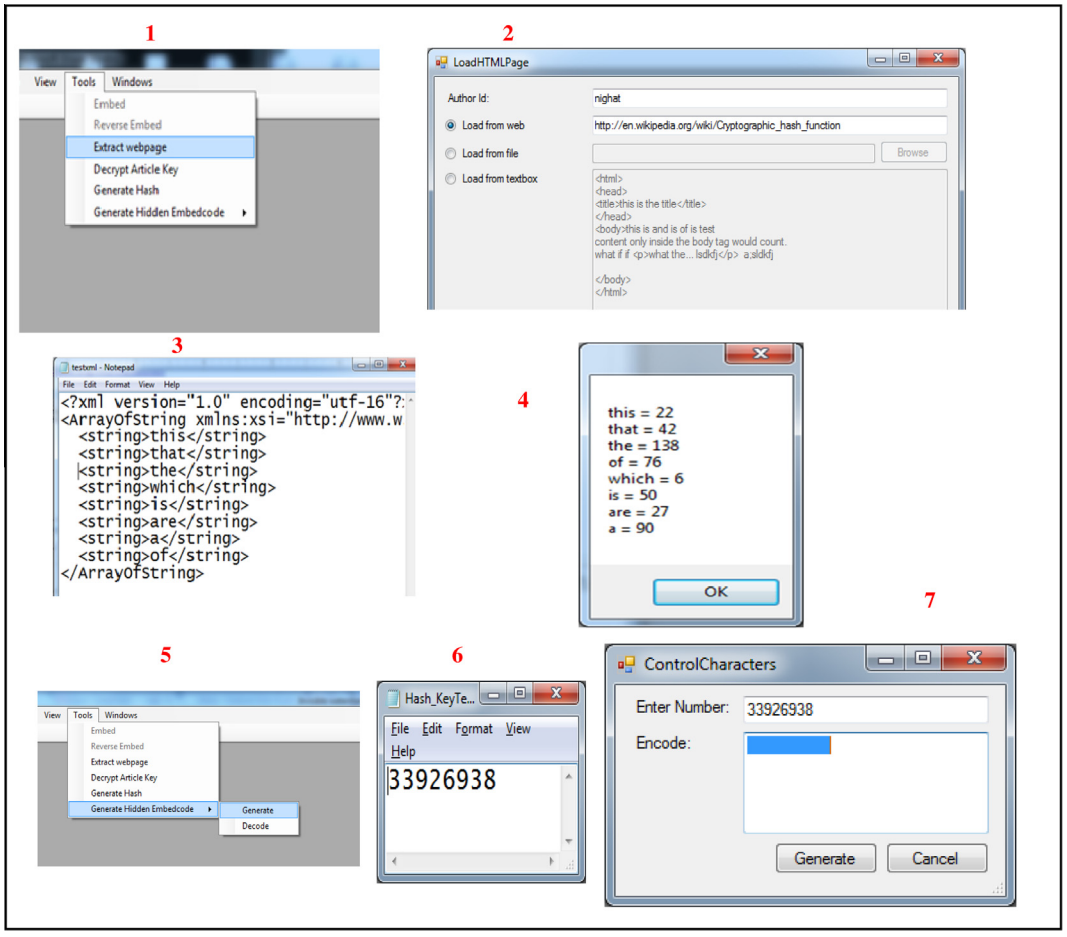

Fig. 3. Experimental steps.

Table 1
Experimental results.

Experimental Results			
URL's	**Watermarks occurrence**	**Encrypted hash value**	**Invisible watermarks**
1. http://en.wikipedia.org/wiki/Digital_watermarking	{is=54,are=12, to=58, of=45}	27041469	□□□□□□□□□□□
2. http://en.wikipedia.org/wiki/Cryptography	{is=95,are=58, to=161, of=258}	22322815	□□□□□□□□□□□
3. http://en.wikipedia.org/wiki/Hash_function	{is=89,are=31, to=107, of=141}	47112568	□□□□□□□□□□□

watermarks e.g. whitespace replacement, line break, color replacement and empty tags, used as a source location for embedding the watermarks. In this research <meta> tag has been used to embed the watermark as it is mainly used for providing information about the program details such as date of creation, author name or any extra information which is not meant to display on the browser. Watermarks can be used as invisible, visible (without converting into whitespaces) and disguised (by converting the watermarks into date or time format) to further enhance this research.

Acknowledgements

This research project is funded and supported by Research Centre Institute of Effat University, Jeddah Kingdom of Saudi Arabia under RCI research grants for year 2011–2012. Dr. Nighat Mir is an Assistant Professor in the department of Computer Science, College of Engineering at Effat University.

Appendix A

1. I/P: Input
2. O/P: Output
3. TW: Text Watermarks
4. FoOcc: Frequency of Occurrence
5. EOF: End of File
6. EN_TW: Encrypted Text Watermarks
7. Cover file: Object to carry the secret message
8. MsgChk: Message Check

9. H_Value: Hash Value
10. IV_W: Invisible Watermarks
11. u200A: Memory Control Character with no face or appearance
12. u202F: Memory Control Character with no face or appearance
13. u205F: Memory Control Character with no face or appearance

References

Arnold, M., Schmucker, M., & Wolthusen, S. D. (2003). Techniques and applications of digital watermarking and content protection. *Artech House Computer Security Series*.

Atallah, M. J., McDonough, C., Nirenburg, S., & Raskin, V. (2000). Natural language processing for information assurance and security: An overview and implementation. In *Proceedings of 9th ACM/SIGSAC new security paradigms workshop*, Cork, Ireland.

Atallah, M. J., Raskin, V., & Crogan, M. (2001). *Natural language watermarking: Design, analysis, and a proof-of-concept implementation [C]. Information hiding*. Berlin: Springer.

Atallah, M., Raskin, V., Hempelmann, C. F., Karahan, M., Sion, R., Topkara, U., et al. (2002). Natural language watermarking and tamperproofing. In *Fifth information hiding workshop*. Springer Verlag.

Benjamin, W. (1969). In H. Arendt (Ed.). *Illuminations*. New York: Schocken Books.

Bo, L., Wei, L., Yuan-Yuan, C., Ying-Zhi, C., & Dong-Dong, J. (2009). HTML integrity authentication based on fragile digital watermarking. *Granular Computing, IEEE*.

Brassil, J. T., Low, S., & Maxemchuk, N. F. (1999). Copyright protection for the electronic distribution of text documents. *Proceedings of the IEEE, 87*(7).

Brassil, J. T., Low, S., Maxemchuk, N. F., & Gorman, L. O. (1995). Hiding information in document images. *Proceedings of the 29th Annual Conference on Information Sciences and Systems*. Johns Hopkins University.

Gerasimov, N. E., Minin, I. V., & Minin, O. V. (2007). Stealthographic protection of intellectual property. In *Www Documents H Scientific Symposium Technomat & Infotel*, Bulgaria.

Katzenbeisser, S., & Petitcolas, F. A. P. (2000). Information hiding: Techniques for steganography and digital watermarking. *Artech House*. ISBN: 1-58053-035-4.

Li, Q.-C., & Dong, Z.-H. (2008). Novel text watermarking algorithm based on Chinese characters structure. *IEEE* 978-0-7695-3498-5/08.

Lu, C.-S. (2005). *Multimedia security: Steganography and digital watermarking techniques for protection of intellectual property*. Taiwan, ROC: Institute of Information Science, Academia Sinica.

Lu, H., DingYi, F., XiaoLin, G., XiaoJiang, C., XinBai, X., & Jin, L. (2009). A new Chinese text digital watermarking for copyright protecting word document. *IEEE* 978–0-7695-3501-2/09.

Martínez, S., Sanchez, D., Valls, A., & Batet, M. (2012). Privacy protection of textual attributes through a semantic-based masking method. *Journal of Information Fusion, 13*, 304–314.

Mir, N., & Hussain, S. A. (2011). Secure web-based communication. *Elsevier. Procedia Computer Science, 3*, 556–562.

Mulwad, V., Li, W., Joshi, A., Finin, T., & Viswanathan, K. (2011). Extracting information about security vulnerabilities from web text. *IEEE* 978-0-7695-4513-4/11.

Renard, L. (1999). Cut and paste 101: Plagiarism and the net. *Educational Leadership, 57*(4).

Topkara, M. (2007). *New designs for improving the efficiency and resilience of natural language watermarking*. PhD Thesis, Purdue University, West Lafayette, Indiana.

Topkara, U., Topkara, M., & Atallah, M. J. (2006). The hiding virtues of ambiguity: Quantifiably resilient watermarking of natural language text through synonym substitutions. In *Proceedings of ACM multimedia and security conference*, Geneva, Switzerland.

Topkara, M., Topraka, U., & Atallah, M. J. (2007). Information hiding through errors: A confusing approach. In *Proceedings of SPIE international conference on security, steganography, and watermarking of multimedia contents IX*, 6505.

Turner, P. A. (1990). COPYCAT: A system for the distribution of copyright cataloging information. *IEEE*.

Zhao, J., & Koch, E. (1995). Embedding robust labels into images for copyright protection in Intellectual Property Rights New Technologies. In *Proc. KnowRigh* (pp. 242–251).

Zhao, X., & Xue, L. (2009). Managing interdependent information security risks. In *Thirtieth international conference on information systems*, Phoenix, Arizona.

Zhou, T., & Li, L. (2009). A secure web-based watermarking scheme for copyright protection. In *Sixth web information systems and applications conference, IEEE*.

Zhu, J., Xu, B., Jiang, B., & Chen, W. (2010). Identifying harmful web pages in laboratory information security management. *IEEE* 978-1-4244-5874-5/10.