



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.

© 2013 Elsevier Ltd. All rights reserved.

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 internet users. 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

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 space-bar, 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

Download English Version:

<https://daneshyari.com/en/article/350720>

Download Persian Version:

<https://daneshyari.com/article/350720>

[Daneshyari.com](https://daneshyari.com)