International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:1, 2011

# Web Page Watermarking: XML files using Synonyms and Acronyms

Nighat Mir, Sayed Afaq Hussain

***Abstract*—**Advent enhancements in the field of computing have increased massive use of web based electronic documents. Current Copyright protection laws are inadequate to prove the ownership for electronic documents and do not provide strong features against copying and manipulating information from the web. This has opened many channels for securing information and significant evolutions have been made in the area of information security. Digital Watermarking has developed into a very dynamic area of research and has addressed challenging issues for digital content. Watermarking can be visible (logos or signatures) and invisible (encoding and decoding). Many visible watermarking techniques have been studied for text documents but there are very few for web based text. XML files are used to trade information on the internet and contain important information. In this paper, two invisible watermarking techniques using Synonyms and Acronyms are proposed for XML files to prove the intellectual ownership and to achieve the security. Analysis is made for different attacks and amount of capacity to be embedded in the XML file is also noticed. A comparative analysis for capacity is also made for both methods. The system has been implemented using C# language and all tests are made practically to get the results.

***Keywords*—**Watermarking, Extensible Markup Language (XML), Synonyms, Acronyms, Copyright protection.

## I. Introduction

With the advancement in telecommunication and computing a rapid growth of electric documents processing on Internet is evolving every day. This perhaps has evolved and matured the concepts of e-business, e-commerce and e-learning.

Electronic publishing has played a significant role in the field of internet technologies and web-development. But this has also evolved the subject of Information Security in recent years very significantly. With the great enhancements in the field of information security the three basic principles (Confidentiality, Integrity and Availability) are doubled and more principles (Possession, Authenticity and Utility) are also added in the security model.

Un-authorized use of text by copying from the internet has become a common practice and has a great effect on the privacy of data. Electronic documents are exposed to various threats like copying, redistributions, destruction, forgery and tampering of data.

Copyright protection is no more enough for the electronic documents as copying and manipulating information is not difficult. Digital Watermarking methods are considered a strong mechanism to identify the original owner and to prove the intellectual property. Watermarking is a branch of information security in which additional ownership information like name, logo, ISBN or signature is added to the content. This can be applied to any digital media like audio, video, image or text to prohibit the un-authorized use and duplication. Various methods have been studied and applied for the multimedia objects but a few for text or electronic text without altering its integrity.

In Digital watermarking a hidden marker is embedded to the data which is generally un-observable and can be only drained by special detector. The main aim of digital watermarking to use human's insensitive perceptual organs and it does not change the basic characteristics. [1]

With the ever increasing growth of internet users all over the world, it is very important to secure the web pages. There is a wide bandwidth present in web pages for information hiding and many robust techniques can be developed for web page watermarking. Web page watermarking is to achieve the integrity of web pages which is a very popular and rich source of information.

HTML and XML are main tools for web development. Even scripting code is also translated by the browser into HTML format at the end. XML files are used to exchange information on internet and are very sensitive for the owners [2]. Due to its sensitivity, importance of XML security is growing everyday and different techniques have been developed for its integrity.

Due to the big amount of data published in the form of XML its protection is becoming an important requirement. Watermarking scheme for XML files should be based on the usability of data and the underlying semantics like key attributes and functional dependencies [3].

## II. Related Work

Qijun Zhao, Hondtao Lu [4] have proposed scheme for the tamper proof web pages in which watermarks are generated on the basis of the Principal Component Analysis (PCA) technique. These watermarks are then embedded in HTML

Nighat Mir is with the Department of Computer Science, College of Engineering, Jeddah, Kingdom of Saudi Arabia. phone:966-2-6364300-2324; fax:966-6377447; email:nmir@effatuniversity.edu.sa.

Dr. Sayed Afaq Hussain is with the Department of Computer Engineering, Riphah International University, Islamabad, Pakistan; email: drafaqh@gmail.com.

tags using upper and lower cases. Shingo, Kyoko [5] have proposed some methods like using empty elements, white spaces in tags, attribute and element ordering to hide information in XML files. These techniques can be used during the construction of XML page to conceal the hidden information. Adnan, Osama [6] have proposed watermarking for electronic documents that contains justified paragraphs and irregular line spacing where spread spectrum technique is used to match the effects. Watermark is embedded by slightly increasing or decreasing the spaces based upon the bit value of watermark.

## III. Proposed Methodology and System Models

XML files are watermarked using synonyms and acronyms. Both types of information can be embedded in one XML file but one technique is applied at a time. An XML file is used as a carrier on the sender side. The file is watermarked using synonyms or acronyms. While embedding a list of synonyms or acronyms can either be taken ready or can be created dynamically. Once watermarked using the synonyms or acronyms the XML file is checked against its validity. Only valid XML files with their standard definitions are accepted by the system. HTML files are used to display the information enclosed in the XML files on the browser. The current system is not restricted with any specific type or version of browser. The system has also a flexibility to create XML files at runtime.

At receiver end, the watermarked information displayed in the browser can only be decoded if there exists a list of synonyms or acronyms even if XML file is breached. This adds a security standard for the authentication and illegal ownership of information or XML files. For the authorized party who has the access to the XML file and has a list will require to compare the XML file with the list. A comparison is performed recursively on the XML file until the end of file and whenever a match is found for a word in the list, it is replaced with its synonym. Same is applied in case of acronyms when a match is found, the acronym is replaced with its manipulated definition.

## IV. Units

XML is taken as an input object to be watermarked on sender side and Tags are constructed according to the manipulated Synonyms List (SL) or Acronyms List (AL), one at a time. XML file is validated and translated by HTML and the relative information is displayed on the browser.

At reverse, displayed HTML information needs to be decoded using the lists (SL or AL). Information cannot be decoded unless having a list of synonyms or acronyms.

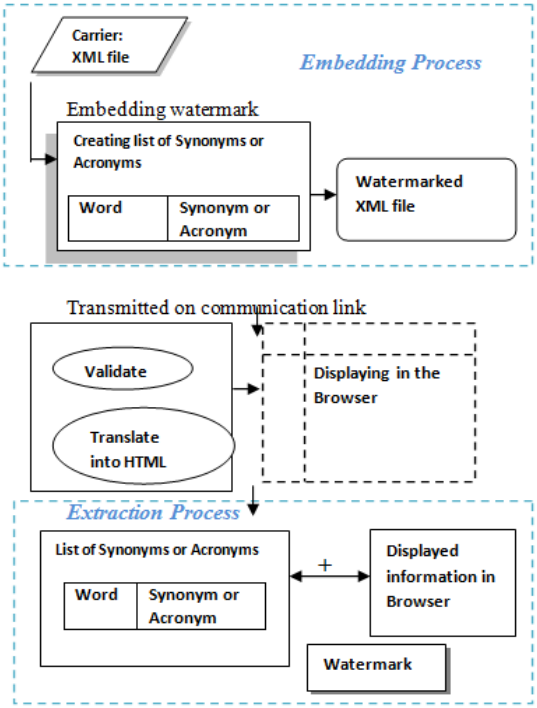


Fig. 1 System Diagram

## V. Watermarking Techniques

### A. Using Synonyms

A Synonyms List (SL) is a database of words with their synonyms and a dictionary of SL is first created. XML file is constructed based upon the manipulated Synonyms List (SL) which acts as a watermark. This watermarked file has information to be exchanged over the Internet and is displayed by HTML on browser. At reverse to decode or to extract watermark, one needs to have the Synonyms List (SL). Once you have the list, XML file is compared with the SL. If there is a match found word is replaced with its synonym otherwise ignored and the process is recursively applied to extract the watermark.

### B. Using Acronyms

An Acronyms List (AL) is a database of acronyms with its definition and a dictionary of AL is first created. All standard and commonly used acronyms are selected while creating an AL. XML file is constructed based upon the manipulated Acronyms List (SL) which acts as a watermark. This watermarked file has information to be exchanged over the Internet and is displayed by HTML on browser. At reverse to decode or to extract watermark, one needs to have the Acronyms List (AL). Once there is a list, XML file is compared with the AL. If there is a match found definition of acronym is replaced with its manipulated definition otherwise

ignored. Therefore, this process is recursively applied to extract the watermark till the end of file.

## VI. Implementation and Results

A C# program is written to test both methods. In each case a list of 100 words with their synonyms and acronyms is created dynamically. The program has a flexibility to use the readymade available list or can create one at the runtime. At reverse, a recursive procedure is applied on the XML file to extract the embedded watermarks by a matching technique. Whenever a matching synonym/acronym is found on the page it is treated according to the mechanism of replacing with its value (value is synonym or a manipulated definition in both cases).

### A. Results Using Synonyms List (SL)

*Original data with watermarks:*

Digital watermarking is the process of embedding information into a digital signal in a way that is difficult to remove. The signal may be audio, picture, or a video. If the signal is copied, then the information is also carried in the copy. A signal may carry several different watermarks at the same time. In visible watermarking, the information is visible in the picture or video [6].

TABLE I
Synonyms List (SL)

| Word(watermark) | Synonyms |
|---|---|
| process | rout |
| information | in order |
| signal | Hint |
| difficult | complex |
| remove | eliminate |
| visible | Evident |

*After replacing with their synonyms:*

Digital watermarking is the rout of embedding in order into a digital hint in a way that is complex to eliminate. The signal may be audio, picture, or a video. If the signal is copied, then the information is also carried in the copy. A signal may carry several different watermarks at the same time. In evident watermarking, the information is visible in the picture or video.

*Watermarks*

*rout, in order, hint, complex, eliminate, evident*

### B. Results Using Acronyms List (AL)

*Original data with watermarks*

Advanced Research Projects Agency Network (ARPANET) for the United States Department of Defense. A computer network allows sharing of resources and information among interconnected devices. Networks are often classified as local area network (LAN), wide area network (WAN), metropolitan area network (MAN), personal area network (PAN), virtual private network (VPN), campus area network (CAN), storage area network (SAN), and others, depending on their scale, scope and purpose, e.g., controller area network (CAN) usage, trust level, and access right often differ between these types of networks [7].

TABLE II
Acronyms List (AL)

| Words (watermark) | Acronyms | Definitions |
|---|---|---|
| advanced research projects agency network | ARPANET | analysis report presents ambiguous notions |
| local area network | LAN | low attenuation navigator |
| wide area network | WAN | wide attenuation navigator |
| metropolitan area network | MAN | medium attenuation navigator |
| personal area network | PAN | parallel attenuation navigator |
| virtual private network | VPN | various attenuation navigator |
| campus area network | CAN | compressed attenuation navigator |
| storage area network | SAN | strong attenuation navigator |

| | | |
|---|---|---|
| controller area network | CAN | complex attenuation navigator |

*C.Analysis*

While testing it has been observed that both methods are unnoticeable and secure against attacks like insertion, modifications and deletion in case of having no access to XML file. Different attacks applied on the XML files and it has been noticed that XML file can exist well with insertion and modification attacks unless the list is not accessible or breached. In case of insertion attack, the inserted words will be ignored while comparing with the list of synonyms/acronym and there are negligible chances of loosing watermarks. In case of modification at bits level, there are enough chances to recover watermarks, but if the whole word is modified (at byte level) then there is a less chance to recover watermarks. However, in case of deletion, there are enough chances to lose watermarks because not every word from the list is supposed to be in the XML file, in case when we are using a big list/database of synonyms/acronyms. But this issue can be resolved by limiting the list of certain numbers of synonyms/acronyms used in a file. In this case the system will not be very robust as we will have a small database but there are less or no chances for deletion attacks. In our experiments both case have been tested and achieved the same results as discussed earlier.

A capacity comparison has also been made on XML files using both techniques and it has been noticed that both methods contain a good capacity. It has been test for both methods and it is noticed that capacity of embedding depends on the numbers of synonyms and acronyms used in a file.

Capacity has been measured at byte level with respect to the numbers of synonyms/acronyms used. A file with three synonyms shows that it has a capacity of 3, which means that three words are replaced in the file. Similar results have been noticed using the acronyms method. 11 different lists/XML files have been created with different numbers of synonyms/acronyms used starting with 3 words replaced then 5, 7, 9, … 25. (3,5,7,9,11,13,15,17,19,21,23,25) and then tested with both synonyms and acronyms techniques for the capacity and it has found that they move sequentially with the increase of numbers and show the same level of capacity (3-synonyms/acronyms-3(capacity), 5-5, 7-7, …, 25-25). This shows a good relation.

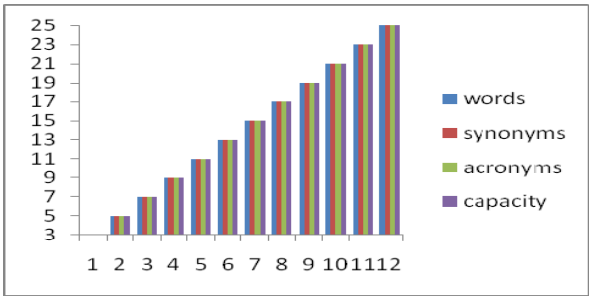


Fig. 2 Capacity Comparison

TABLE III
CAPACITY FOR SYNONYMS AND ACRONYMS

| Word (watermark) | Synonyms | Acronyms | Capacity |
|---|---|---|---|
| 3 | 3 | 3 | 3 |
| 5 | 5 | 5 | 5 |
| 7 | 7 | 7 | 7 |
| 9 | 9 | 9 | 9 |
| 11 | 11 | 11 | 11 |
| 13 | 13 | 13 | 13 |
| 15 | 15 | 15 | 15 |
| 17 | 17 | 17 | 17 |
| 19 | 19 | 19 | 19 |
| 21 | 21 | 21 | 21 |
| 23 | 23 | 23 | 23 |
| 25 | 25 | 25 | 25 |

## VII. CONCLUSION AND FUTURE WORK

In this paper, we have presented two invisible watermarking techniques using Synonyms and Acronyms for web based information contained in XML files. Both methods have been applied on XML files and these XML files are tested with a static website developed using HTML. These techniques have been implemented in C# language using .net framework and analyzed with respect to some of the variables like security, appearance, perceptibility and bandwidth.

In this research, Synonyms and Acronyms techniques are applied on text information contained in XML file. The idea can be extended towards the other types of data as well, as XML may contain different types of database queries, video, audio or images. Moreover, these techniques can also be applied directly on HTML files, applying on HTML files will not be as secure as XML as they are more accessible with their source information but with a high capacity as web pages contain a good amount of bandwidth.

## REFERENCES

[1] Gengming Zhu, and Nong Sang, "Watermarking Algorithm Research and Implementation Based on DCT Block", World Academy of Science, Engineering and Technology 45 2008
[2] Xuan ZHOU, HweeHwa PANG, Kian-Lee TAN, Dhruv MANGLA, "WmXML: A System for Watermarking XML Data", Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005
[3] Mohammad Laheen, Sun XingMing, "Techniques with Statistics for Web page Watermarking" 2005, NSFC No.60373062.
[4] Qijun Zhao, Hondtao Lu, "PCA-based web page watermarking", Elsevier Science Inc., Vol. 4, 2007
[5] Shingo, Kyoko, Ichiro, Osamu, "A Proposal on Information Hiding Methods using XML", Mitsubishi Research Institute, Communication Research Laboratory, Yokohama National University and The University of Tokyo.
[6] http://en.wikipedia.org/wiki/Digital_watermarking
[7] http://en.wikipedia.org/wiki/Computer_network