Securing web contents through invisible text watermarking for copyright protection

[1]Sameeka Saini, [2]Kalpesh Prajapati

[1]ME(CSE), [2]ME(IT)

***Abstract*— In the growing era internet has become a wide source of information. The ease of accessing the information and distribution of information over the internet has leads to many threats on information security. The contents and documents must be protected from unauthorized access, distribution and forgery. This can be achieved through watermarking. In this paper a robust and invisible technique is used to generate the watermark and then secured it using cryptographic hash function and finally making it invisible.**

***Index Terms*— Digital Watermarking, Steganography, SALT, Cryptography, XML, Copyright, HASH.**

I. INTRODUCTION

The rapid growth of the Internet usage and the dependency on internet needs security against various malicious activities such as fraud, copyright, online vulnerabilities, etc. Ease of search and access huge amount of knowledge online has become a part of everyday life. The original author or owner doesn't have control on how their data is been used or distributed over the internet. Copying of information from internet has become a regular activity. Securing this information is necessary and for the same there are many algorithm and techniques, some of which includes Cryptography, Steganography and Watermarking [1].

Cryptography is a technique of making the text or the message into unreadable format by using some permutation and combination on it. The process is commonly known as encryption and the reverse process is known as decryption. Steganography is a technique of hiding the secret information on some carrier file that can be anything i.e. image, audio or video. Digital Watermarking is a technique of embedding some secret information in the main digital content to provide security, integrity and authentication.

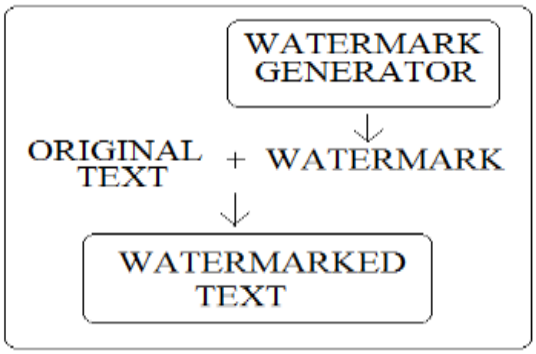

Figure 1: The process of watermarking

The process of embedding and extracting a watermark to and from a digital text document which uniquely defines the original owner of it is known as Digital Text watermarking. A watermark is basically used for identifying either the originator or authorized user. Fig 1 shows the process of watermarking. The Digital Watermarking is the best technique for copyright protection of electronic media [2].

II. LITERATURE REVIEW

Nighat Mir, sayed A. Hussain [1] proposed different watermark embedding techniques and has shown the comparative analysis of them. Table 1 shows the comparative analysis of those techniques. From the above comparison the synonyms and acronyms method seems to be the best for embedding the watermark in digital content but the list of synonyms and acronyms is must or this type and if the list is known to the attacker the scheme is no more secure.

Table 1: Comparison of different embedding methods.

Methods	BLINDNESS	SECURITY	CAPACITY	ROBUSTNESS
White space	Yes	YES	Medium	More
White space	No	YES	Medium	More

Empty tag	Yes	YES	Low	More
Random Character	No	YES	High	Less
Color replacement	Yes	YES	Low	More
Word space	Yes	YES	Medium	More
Synonyms & Acronyms	Yes	YES	High	More

Nighat Mir [2] proposed a novel robust web text watermarking algorithm. He calculated no of occurrences of words based on syntax and semantic rules and applied HASH on it to generate an invisible watermark. This watermark is made invisible by no face control characters and embedded it in meta tag of HTML page.

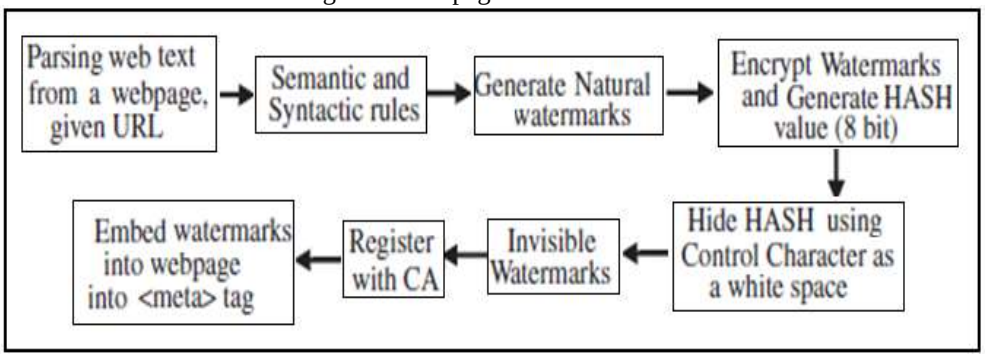

Figure 2: Watermark generation, conversion & embedding.

The technique is implemented on HTML and can be applied on other languages and the different embedding techniques can also be used in place of no space control characters.

Makarand, Nitin & J.B [3] proposed a new technique for text watermarking using natural languages grammatical rules. They have taken first hundred frequently occurring words along with author id. This is then encrypted using AES to get a key of arbitrary length.

Jaiswal Raj & Nitin Patil [4] proposed a technique for web document protection using Unicode. They generated watermark by two techniques i.e Unicode conversion and HTML text coding. They took random text and converted it into Unicode and then to HEX. A table is created for every HEX digit there is a tag. The corresponding digits of HEX is converted to HTML tag Coding and then this tag coding is embedded into the web page.

Li Bo, Li Wei, Chen Y., Jiang D. D, Cui Y. [5] in their paper showed the comparison of different hash algorithm based on the performance. Following table shows the comparison and it is concluded that SHA-1 requires less operation time.

Table 2: The performance of various Hashing comparison.

Algorithm/Attribute	Hash Value	OPERATION TIME (MS)
MD2	128bits	46
MD5	128bits	47
SHA-1	160bits	16
SHA-256	256bits	16
SHA-348	384bits	31
SHA-512	512bits	32

The paper also compares the seven information hiding methods based on HTML. The techniques are based on the parameter of capacity. The aim was to ensure the integrity of the webpage, not needed too much information. For XML web pages there are different techniques. Nighat Mir, Sayed Afaq hussain [6] proposed a web page watermarking scheme for XML files using Synonyms & Acronyms.

T. Romaric/E. Damiani & N. Bennani [7] proposed a technique in which a locator selector algorithm selects the position or tag where the watermark is to be embedded. They used fuzzy query language that uses fuzzy predicates to retrieve information from a XML dataset. As future work they planned to produce a complete instantiation of their proposed work.

T. Romaric, S. Cimato & N Bennani [8] proposed a technique that was against the XML signature attack. Their solution consists of an embedding and a verification procedure. Embedding consists of three steps. The first one is the expression of constraints on the structure of the XML file using Schematron language. The second step is the computing of absolute coordinates of all the nodes in the XML file. In the third step, identify which coordinates have to be watermarked, in order to successively perform the verification phase. The tags where watermark is embedded are those, which present a high usability degree. As future work they mentioned it to enlarge the set of constraints that could be taken into account and make the XML file as flexible as possible.

Jesus U., Q-Torrero & G. Erickson [9] compared four main techniques for watermarking XML i.e WmXML, XML Streams watermarking, Natural language watermarking & selective watermarking approach. WmXML watermarking method is applied to simple, uncompressed XML files and consists of relatively simple algorithms and implementation methods for watermarking data in XML documents.

III. PROPOSED WORK

The existing work in [2] can be made more secure. Since the watermark is generated using the source code and counting the frequency of the words, if the algorithm and XML file of words is hacked by the attacker the scheme remains no more secure. In proposed work based on grammatical rules the watermark is generated and then it is secured using Cryptographic hash function SHA-1 and then it is made invisible and finally embedded into META tag of source code. The concept of SALT is used to make the scheme more secure. The following figure shows the process of proposed work.

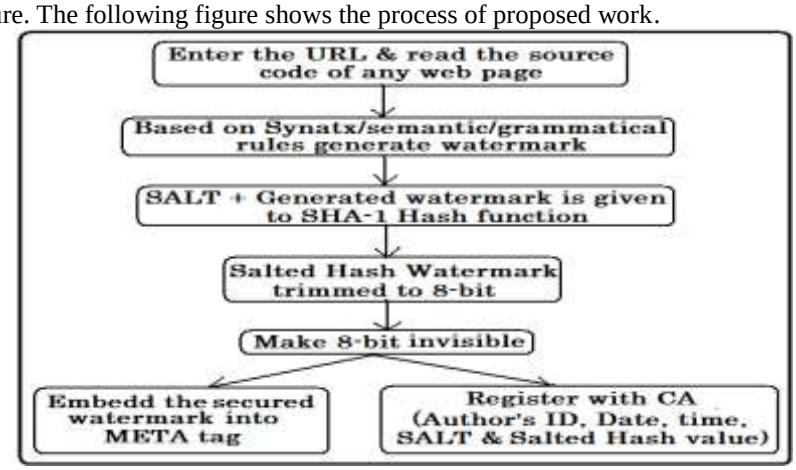

Figure 3: Watermark generation and securing process.

The Use of SALT adds security in the work. Since the SALT and its position is not fixed and only known to the author the brute force attack is impossible for attacker. Also for others attack such as insertion, deletion and modification attack this technique is more secured and more robust.

IV. RESULT ANALYSIS

Following figures are the snapshot of the entire proposed process.

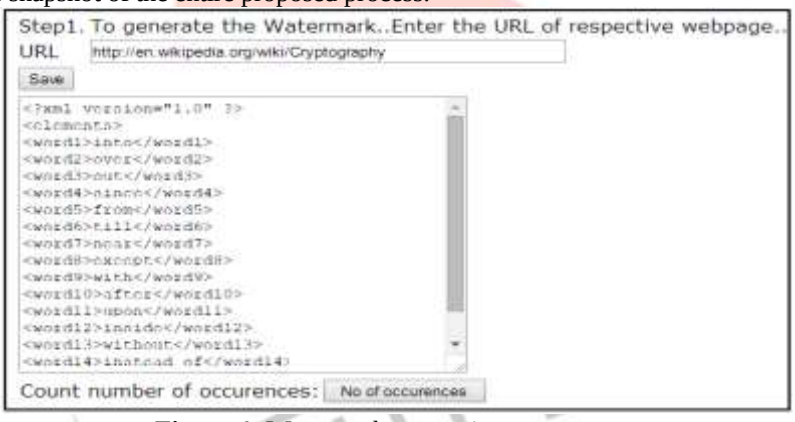

Figure 4: Watermark generation process.

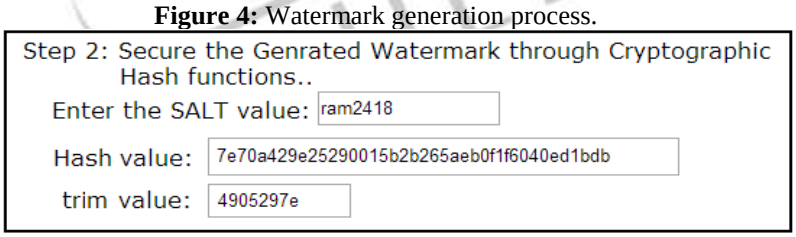

Figure 5: Securing generated watermark process.

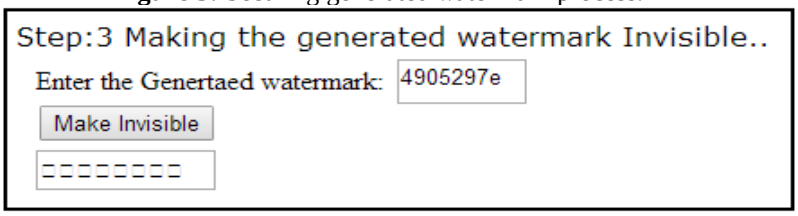

Figure 6: Making watermark invisible.

The proposed work is tested on 3 websites and the following table shows the details of results obtained.

Table 3: Experimental result of proposed work.

URL's	Watermark occurrence	SALT value	Encrypted hash value	Invisible watermark

http://en.wikipedia.org/wiki/Digital_watermarking	221190009500001	12ab34	22cb6037	========
http://en.wikipedia.org/wiki/Cryptography	589163800251910501	Ram2418	4905297e	========
http://en.wikipedia.org/wiki/Hash_function	121222802053200 4	2432107	2811f05f	========

The following are the results of testing of proposed algorithm in XML file.

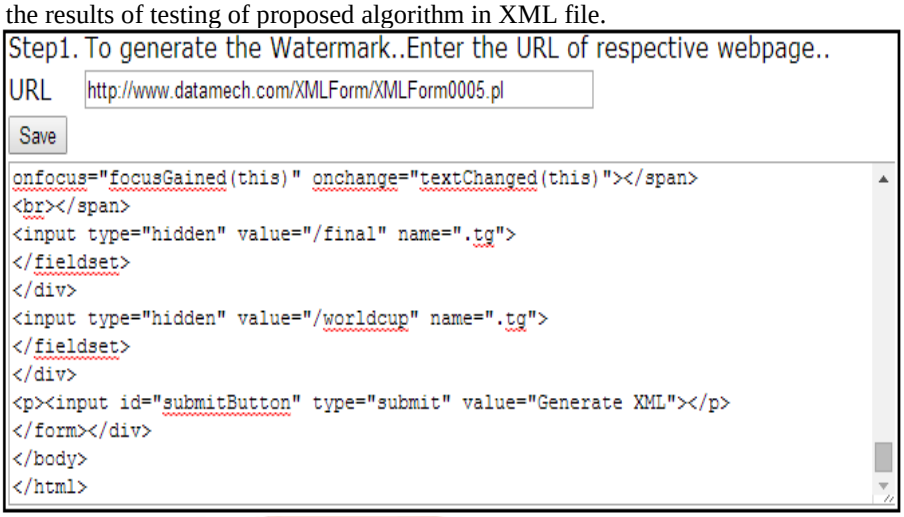

Figure 7: Taking XML file as input file.

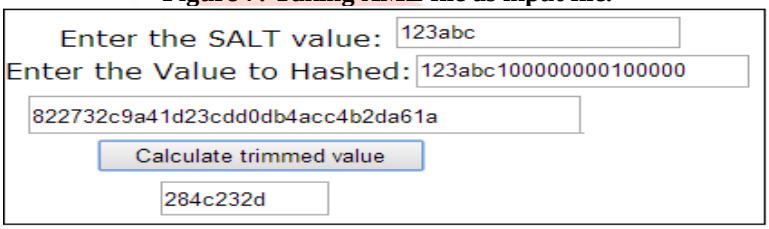

Figure 8: Calculating Hash value.

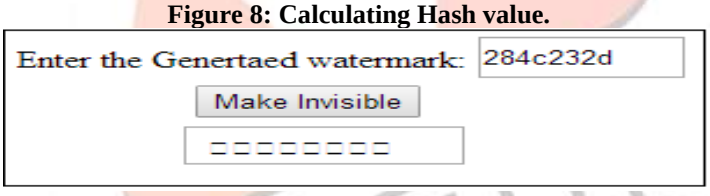

Figure 9: Making Hash value invisible.

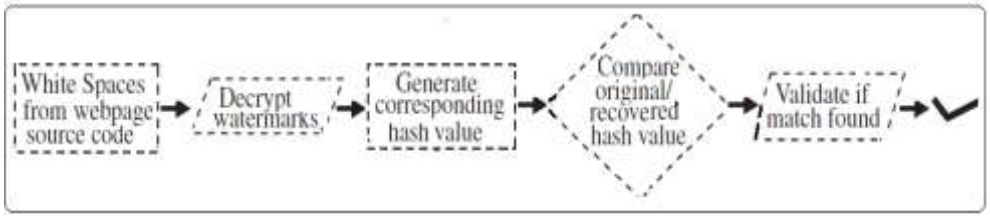

Figure 10: Validation process.

The following graph shows that SHA-1 takes less processing time and hence it is used in this technique.

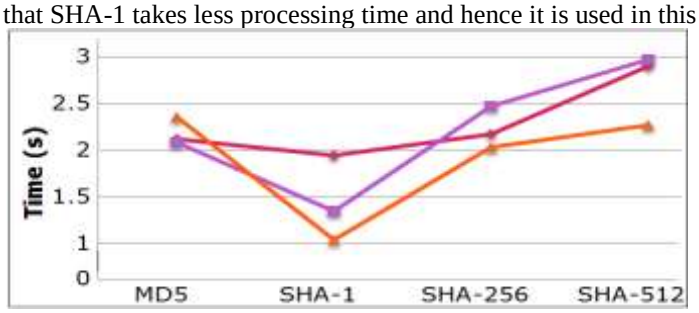

Figure 7: Graph showing timing comparison between MD5, SHA-1, SHA-256 & SHA-512.

The following figure shows the effect of adding the watermark on the webpage.

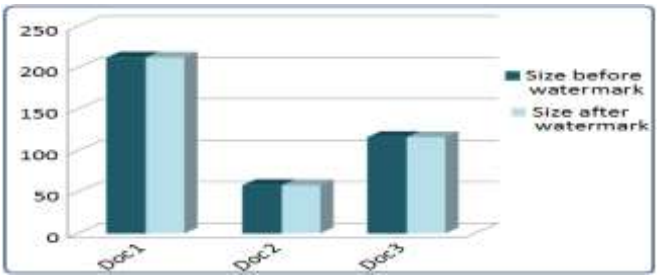

Figure 8: Graph showing Effect of adding watermark on the size of webpage.

V. CONCLUSION AND FUTURE WORK

Digital watermarking is a copyright protection technique used to embed specific data in a cover file to prevent illegal use. The issues in the existing work were regarding security of the generated watermark along with improvised authentication process of watermarking. By using the SALT along with cryptographic Hash function has increased the security of the algorithm and has made it secure against brute force attack also. The same algorithm is tested on XML files also and is more secure and more robust on them also. The technique doesn't consider the images, video and PDF present in the webpage, this work can be done as a future work.

VI. ACKNOWLEDGMENT

We, Sameeka Saini and Kalpesh Prajapati would like to acknowledge and extend my sincere gratitude to our Guide Asst. Prof. Mohammed Hussain Bohara for their regular guidance and encouragement in all the aspects for successfully completion of our research work.

REFERENCES

[1] Mir, N., Hussain, S.A (2011), "Secure web-based communication," Elsevier, Procedia Computer Science, 3, 556-562

[2] Nigat Mir (2014), "Copyright for web content using invisible text watermarking," Elsevier, Computers in Human behavior, 648-653.

[3] Makarand L. Mali, Nitin patil, J.B patil (2013), "Implementation of text watermarking technique using natural language watermarks," International Conference on communication systems and network technologies, 482-486.

[4] Jaiswal Raj, Patil Nitin, "Implementation of a new technique for web document protection using Unicode"

[5] Bo, L., Yuan-yuan, C., & Dong-Dong, J. (2009), "HTML integrity authentication based on fragile digital watermarking," Granular Computing IEEE.

[6] Nighat Mir, Sayed afaq hussain (2011), "Web page watermarking: XML files using Synonym & Acronyms," World academy of science, engineering & technology.

[7] Tchokpon Romaric/Ernesto Damiani & Nadia bennani (2012), "Robust XML watermarking using fuzzy Quries," IEEE 36th International conference on computer software and applications workshops, 433-438.

[8] Tchokpon Romaric, Stelvio Cimato & Nadia bennani (2012), "Ensuring XML integrity Using watermarking techniques," IEEE 8th International conference on signal image technology and internet based systems, 668-674.

[9] Jesus Ubaldo Quevedo-Torrero & Geno Erickson (2012), "Watermarking XML structures Using Multi-valued dependencies," IEEE 9th International conference on information technology, 554-559.

[10] Zhou, X, Pang, H, Tan, K, Mangla, D. WmXML, "A System for Watermarking XML Data," Proceeding of the 31st VLDB Conference, 2005.

[11] Jalil Z. Farooq, Zafar H., Sabir M., and Ashraf E.(2010), "Improved zero text watermarking algorithm against meaning preserving attacks," World academy of Science, Engineering & Technology. 540-544.

[12] Turner, P. A. (1990), "COPYCAT: A system for the distribution of copyright cataloging information," IEEE.