

Text Watermarking for Text Document Copyright Protection

Computer Science 725 Term paper

Author: ChaoLi Ou, ID: 2133886, cou002@ec.auckland.ac.nz

Date: 9 June 2003

Abstract:

Text watermarking is an approach for text document copyright protection. Watermarking ensures that a text document carries a secret message containing copyright information so that copyright infringement can be recognized. Three examples of text watermarking – Open Space, syntactic, and semantic-- are studied. These correspond to three text formats: Window's Word, HTML (XML), and LaTeX, and demonstrate the concept in relation to copyright protection.

1 Introduction

The popularity of computers, the internet and photocopy machines has made text document copying easy and without loss of integrity. The copyright of text documents are infringed by illegal copy and distribution. Watermark is one of the approaches for text document copyright protection. Watermark is used commonly in monetary currencies to make counterfeiting more difficult. A watermark can be a variety of objects such as an image, or some copyright information. A watermark is some data that can be extracted later as a proof of the copyright of the document in the situation of digital document.

Text watermarking is a process to embed a watermark into text document. The watermarks can be divided into two types: visible and invisible, and is associated with human vision. The watermark discussed in this paper is invisible. It is some data embedded into text document that can pass easily from one copy to another copy. Therefore, the watermark works like a token that allows the copyright owner to detect illegal copies and prosecute the copyright violator who may be the seller or the owner of the illegal copy.

If a text document is a long document, and then the copyright protection system assigns a secret key to copyright owner. The system keeps the record of secret key, and even a copy of the document. A watermark embeds into text document by the secret key; therefore the secret key can be used to extract the watermark from the text document. The copyright owner may have one or more secret keys, but he had to keep recording of which key for which text document.

There are not many ways to embed a watermark in to a text document except three common methods: open space, syntactic and semantic methods.

Open space methods hide data by adjustment the white space, i.e. the extra white space (between lines or between words) is encoded as 1, and normal white space is encoded as 0.

Syntactic methods hide data by rearranging the order of words in sentences. Some punctuation can be modified without change the sentence meaning. For example, both “You and I” and “I and you” are correct usage. The first usage is encoded as 1, and second usage is encoded as 0. A watermark can be embedded in a text by altering “the usage 1” or “usage 2”.

Semantic methods are hiding data by substituting words with synonyms. The sentence with synonyms substitution is “1” and the original sentence is “0”. When you embed watermark “1”, you choose the sentence with synonyms substitution, otherwise choose the original sentence to embedded watermark “0” into the document.

Document formats discussed in this paper are Window’s word, HTML, and LaTeX which are the most common text format currently used. The aim of this paper is to overview the text watermarking and its three approaches to text document copyright protection. The rest of the paper is as follows. Section 2 describes the text watermarking algorithm. Section 3 gives an example of open space method, syntactic method and semantic method. Section 4 discusses the features of text watermarking. Section 5 concludes and describes future work.

2 Text watermarking algorithm

The text watermarking algorithm consists of four parts: the watermark, the encoder (insertion algorithm), the detector and the comparator (verification or extraction or detection algorithm) [6].

It assumes an original text document O , a secret key $K=k_1, k_2, \dots, k_i$, watermark M and the watermarked text document W . Watermark M is embedded following the key. The insertion function E generates a watermarked text document T' correspond to the input of O , M and K . The function E is represented by

$$E(O, K, M) = W \quad \langle 1 \rangle$$

The detector function D takes a text document H (H is a suspected illegal text document of O . H is at least looks like O) and copyright owner’s key $K=k_1, k_2, \dots, k_i$ then it extract watermark M' . The function D is represented by

$$D(H, K) = M' \quad \langle 2 \rangle$$

A compare function C takes M' as an input to compare with all M recorded in its system data base.

$$C(M', M) = 1; \text{ if } M' = M, \text{ otherwise } C(K', K) = 0 \text{ if } M' \neq M.$$

The following figure illustrates above three functions and the whole text watermarking algorithm. Assume it is in the digital library situation [4].

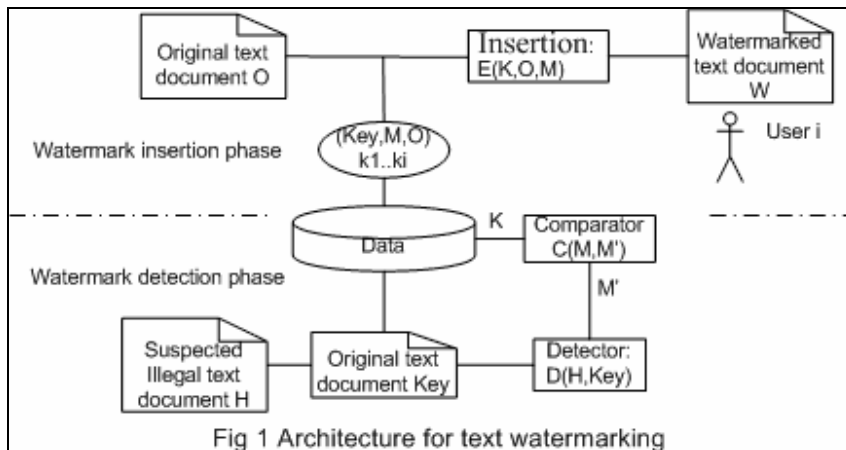


Fig 1 Architecture for text watermarking

The text watermarking algorithm divided into two phases: the watermark insertion phase and watermark detection phase. The insertion function can be implied by difference methods: open space method, syntactic method or semantic method. The key is optional because when the text document is short, no room for embedding key. The watermark M depends on the bit rate of text document. If the text document O is a short text document, then the text document provides a low bit rate, therefore the watermark is a low bit rate object. In detection phase, the system may require to keep original document O for extraction the watermark. When the system keeps many original documents, it is waste of storage space and puts an extra work load on system management. Ideally, the system should not need to keep original documents O ; however, this depends on the insertion method. Open space and syntactic method had to recode original documents O . If a secret key used to insert the watermark, then only part of the document O keeps in the system. The semantic method does not require the original document O to extract the watermark. A great advance for the semantic method is that the system doesn't require the original document O to extract the watermark M .

3. Open space method, syntactic method and semantic method in text watermarking

The example text document is a short document; therefore there is not a secret key for watermark embedding and the water embed begins with the first sentences. It assumes one sentence embeds one bit. The text document example contains 7 sentences, and then the watermark contains 7 bits. We can use this watermark as a look up table entry to put extra information to the watermark. The example text document embeds a watermark by the three methods in three text formats.

3.1 Example of Open space method.

3.1.1 A Microsoft word text document example

A watermark embeds in a section of my summer project report by open space method. A text unit is a sentence end with exclamation marks, question marks, or full stop. The

sentences same with that of original text document is “0”, otherwise is “1”. The sentences and the bit represent of the original document are shown in follow table.

| text_unit | Text Unit (document O) | Bit Value |
|-----------|--|-----------|
| 0 | In this paper I address the problem of quantification the speed and direction of the bacteria colonization. | 0 |
| 1 | It is easy to observe bacteria colonization by comparing the bacteria areas using images taken over time using the same sample. | 0 |
| 2 | However, you can not answer questions, such as: “How fast is the bacteria colonization? What is the direction of the bacteria colonization?” | 0 |
| 3 | I am introducing a method to describe the speed and direction of bacteria colonization by analysis a set of bacteria colonization images. | 0 |
| 4 | The first step is implementing shading correction. | 0 |
| 5 | The second step is bininzation of the images under a specific threshold. | 0 |
| 6 | The third step is counting the pixels in the binary area and finding the radial length of binary area in 8 directions. | 0 |

(Table 1. Original text document O and the initialise bit represents)

The original document O is [000 0000]. In order to embed watermark [101 1001] in the text, I change the gap between the words in the text unit: text_unit[0], text_unit[2], text_unit[3], and text_unit[6]. In the following table, the gap between the underline works is 2 spaces instead of original 1 space; the underline is put for indicate where I put extra space :

| text_unit | Text Unit (document W) | Bit Value |
|-----------|--|-----------|
| 0 | In this <u>paper</u> <u>I</u> address the problem of quantification the speed and direction of the bacteria colonization. | 1 |
| 1 | It is easy to observe bacteria colonization by comparing the bacteria areas using images taken over time using the same sample. | 0 |
| 2 | However, you can <u>not</u> <u>answer</u> questions, such as: “How fast is the bacteria colonization? What is the direction of the bacteria colonization?” | 1 |
| 3 | I am introducing a method to describe the speed and direction of bacteria colonization by analysis <u>a</u> <u>set</u> of bacteria colonization images. | 1 |
| 4 | The first step is implementing shading correction. | 0 |
| 5 | The second step is bininzation of the images under a specific threshold. | 0 |
| 6 | The third step <u>is</u> <u>counting</u> the pixels in the binary area and finding the radial length of binary area in 8 directions. | 1 |

Table 2, the watermarked document W in Microsoft word and the correspond bit value

The text document above with gaps modification is a watermarked document W, and will send to the user_i. The watermark is [101 1001] , and it is integer 89. If we use water mark 89 as a lookup table entry, we can put extra watermark information such as date and name in the lookup table.

3.1.2 A HTML word text document example

Above text document is in html format. The original text is same as Table 1. The watermark [101 1001] is embedded in the text document. The gap between the words in text_unit[0], text_unit[2], text_unit[3], and text_unit[6] are changed. The result is below:

| text_unit | Text Unit (document W) | Bit Value |
|-----------|---|-----------|
| 0 | In this <u>paper</u> I address the problem of quantification the speed and direction of the bacteria colonization. | 1 |
| 1 | It is easy to observe bacteria colonization by comparing the bacteria areas using images taken over time using the same sample. | 0 |
| 2 | However, you can <u>not</u> answer questions, such as: “How fast is the bacteria colonization? What is the direction of the bacteria colonization?” | 1 |
| 3 | I am introducing a method to describe the speed and direction of bacteria colonization by analysis <u>a</u> set of bacteria colonization images. | 1 |
| 4 | The first step is implementing shading correction. | 0 |
| 5 | The second step is bininzation of the images under a specific threshold. | 0 |
| 6 | The third step <u>is</u> counting the pixels in the binary area and finding the radial length of binary area in 8 directions. | 1 |

Table 3, the watermarked document W in html format

The token () is inserted between the words for expanding gap from one space to two space in the browser. The value “1” corresponds to extra space and “0” is normal space. Following figure shows the sentence embedded the code “1”

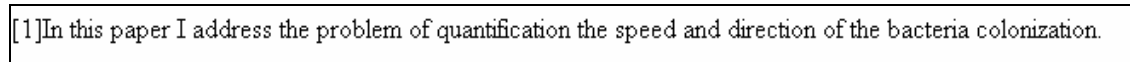


Fig 2 the browser (Microsoft explore 6.0) of the original sentence.

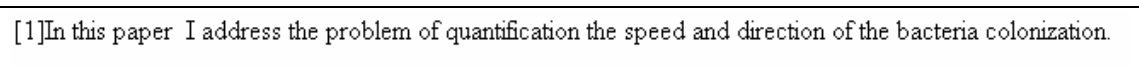


Fig 3 the browser (Microsoft explore 6.0) of W, the code “1” is embedded into sentence by increase the gap between “paper I”.

It is difficult for a normal user to realise that the gap has been modified.

3.1.3 A Latex text document example

The example text document is Latex format. The original text is same with Table 1. The watermark [101 1001] embedded in the text document. The gap between the words in text_unit[0], text_unit[2], text_unit[3], and text_unit[6] are changed. The result is below:

| text_unit | Text Unit (document W.tex) | Bit Value |
|-----------|---|-----------|
| 0 | In this paper\hspace{1cm}I address the problem of quantification the speed and direction of the bacteria colonization. | 1 |
| 1 | It is easy to observe bacteria colonization by comparing the bacteria areas using images taken over time using the same sample. | 0 |
| 2 | However, you can not\hspace{1cm}answer questions, such as: “How fast is the bacteria colonization? What is the direction of the bacteria colonization?” | 1 |
| 3 | I am introducing a method to describe the speed and direction of bacteria colonization by analysis a\hspace{1cm}set of bacteria colonization images. | 1 |
| 4 | The first step is implementing shading correction. | 0 |
| 5 | The second step is binization of the images under a specific threshold. | 0 |
| 6 | The third step is\hspace{1cm}counting the pixels in the binary area and finding the radial length of binary area in 8 directions. | 1 |

Table 4, the watermarked document W in Latex tex format

The token (`\hspace{1cm}`) is inserted between the words for expanding gap from one space to two spaces in the dvi document. A dvi (Device independent file) is generated after compile Latex input file .tex. The dvi can be view by a dvi viewer or similar application. Following figure shows the sentence embedded the code [101 1001], viewed by dvi viewer. An underline is added for showing the extra spacing inserted.

[1]In this paper I address the problem of quantification the speed and direction of the bacteria colonization. [2]It is easy to observe bacteria colonization by comparing the bacteria areas using images taken over time using the same sample. [3]However, you can not answer questions , such as:“How fast is the bacteria colonization? what is the direction of the bacteria colonization? ” [4]I am introducing a method to describe the speed and direction of bacteria colonization by analysis a set of bacteria colonization images. [5]The first step is implementing shading correction. [6]The second step is binization of the images under a specific threshold. [7]The third step is counting the pixels in the binary area and finding the radial length of binary area in 8 directions.

Fig 4, the watermarked document W in Latex document, the underline is added for illustration the extra spacing.

3.1.4 The detection function for open space method

We extract the watermark from document H. The text document is as short as the example. There is not bit rate for insert extra information except the watermark, therefore the protect system had to keep the whole document O for extract watermark M later. First, we compare the H with document O which is store in the protect system string to string to extra watermark M' by finding which sentence has extra space, and then compare M' with M to confirm the watermark. In above example, document H compare with document O, any sentence has extra space between the words is "1", otherwise is "0". After comparing all 7 sentences, have a 7 bit code [1011001].

It is infeasible to keep all original documents O in the system, and to access to all the original documents O to extract the watermark. When the original documents O are long documents, the system may need to keep part of the original document O following a secret key k. For example if a document with 100 sentences, the watermark M is 32 bit. A secret key k indicates which 32 sentences of the document used to embed watermark M. The system only needs keep recoding of these 32 sentences to extract the watermark M from document H.

Note that the original document O and secret key k had to keep safely in the system, otherwise someone will take the original document O or secret key k to embed his own watermark [3][4]. Too many original documents O give the protection system a problem to maintain them.

It is impossible to convert a suspected illegal document H.dvi documents back to H.tex document. The string to string comparison method doesn't work in Latex. The protection system scans H.dvi document uses image process method to build a document profile P'. P' compares with original document profile P to extract H.dvi's embedded watermarking [2].

3.2 Example of syntactic method.

A watermarking is embedded in the example document by modification the syntactic structures of the document. The algorithm is same for LaTeX, GTML, and Window's words format. It is noted that not all sentences in the document can be modification and still has correction grammar and semantic. First, the system goes through whole document to find the amount of sentences can be modified. We use same example document in Section 3.1. The results are shown below:

| text_unit | Text Unit (document O) | Bit Value |
|-----------|--|-----------|
| 0 | I address the problem of quantification the speed and direction of the bacteria colonization in this paper | 1 |
| 1 | It is easy to observe bacteria colonization by comparing the | -- |

| | | |
|---|---|----|
| | bacteria areas using images taken over time using the same sample. | |
| 2 | However, you can not answer questions such as: “How fast is the bacteria colonization? What is the direction of the bacteria colonization?” | 1 |
| 3 | I am introducing a method to describe the direction and speed of bacteria colonization by analysis a set of bacteria colonization images. | 1 |
| 4 | The first step, shading correction is implementing. | 1 |
| 5 | The second step is bininzation of the images under a specific threshold. | -- |
| 6 | The third step is counting the pixels in the binary area and finding the radial length of binary area in eight directions. | 1 |

Table 5, bit value ‘1’ of the original document O by syntactic method

The document is only available for embedding a 5 bit watermark. The text_unit[1] and the text_unit[5] always “0”. The watermark location [1] and [5] is “0”. The protection system can embed watermark [101 1001], the result shows following:

| Bit location | Text Unit (document W) | Bit Value |
|--------------|---|-----------|
| 0 | I address the problem of quantification the speed and direction of the bacteria colonization in this paper | 1 |
| 1 | It is easy to observe bacteria colonization by comparing the bacteria areas using images taken over time using the same sample. | 0 |
| 2 | However, you can not answer questions such as: “How fast is the bacteria colonization? What is the direction of the bacteria colonization?” | 1 |
| 3 | I am introducing a method to describe the direction and speed of bacteria colonization by analysis a set of bacteria colonization images. | 1 |
| 4 | The first step is implementing shading correction. | 0 |
| 5 | The second step is bininzation of the images under a specific threshold. | 0 |
| 6 | The third step is counting the pixels in the binary area and finding the radial length of binary area in eight directions. | 1 |

Table 6, document W embedded watermark [101 1001] by syntactic method

If a suspected illegal document H is html or Microsoft word format, then the protection system does string comparisons with document O to extract the embedded watermarking Key_i. If document H is Latex format, the protection system scans H.dvi, then do image processing, comparisons sentence texture with document O to extract the embedded watermark Key_i. The detection function algorithm is same as that in Section 3.1.4.

3.3 Example of semantic method

The watermarking is embedded in the example document by modifying the sentences with synonym substitution (replace a word with another same meaning word). The algorithm is same for Latex, html, and window words format. The watermark [101 1001] is embedded, and the result is below:

| Bit location | Text Unit (document W) | Bit Value |
|--------------|--|-----------|
| 0 | In this paper I deal with the problem of quantification the speed and direction of the bacteria colonization. | 1 |
| 1 | It is easy to observe bacteria colonization by comparing the bacteria areas using images taken over time using the same sample. | 0 |
| 2 | However, you can not answer questions, such as: “How quick is the bacteria colonization? What is the direction of the bacteria colonization?” | 1 |
| 3 | I am introducing a way to describe the speed and direction of bacteria colonization by analysis a set of bacteria colonization images. | 1 |
| 4 | The first step is implementing shading correction. | 0 |
| 5 | The second step is bininzation of the images under a specific threshold. | 0 |
| 6 | The third step is adding up the pixels in the binary area and finding the radial length of binary area in 8 directions. | 1 |

The watermark detection function is the same with the syntactic method if we keep original document O in the protection system. However, semantic method can extract the watermark without original document O; instead the protection system keeps the hash value HV of sentence’s sum ASCII values. The protection system recodes all sentences’ hash value; or part of the hash value by the secret key for long document. The protection system compares the hash value HV of the document O with the hash value HV’ of the document H to extract H’s watermark M. [1]


Raskin et al. [5] proposed a semantic method by building sentences into text-meaning representations (TMRs) trees, and then the watermark embedded into TMRs trees. This method enlarges the watermark bit rate by converting one sentence into several sentences with same TMRs tree structure. The article does not mention how to conversion sentences to TMRs tree by program. According to the algorithm mentioned in the article, the system had to have a data base to keep all possible TMRs tree and match the tree to the specifics sentence should required human interval.

4. Features of text watermarking in three approaches: Open space method, syntactic method and semantic method

In this section I briefly examine the text watermarking features which would apply to text document copyright protection schema.

The watermarking (M) should extract from watermarked document W reliably. This feature achieved by careful program string comparing or texture comparing. As illustrated in the section 3 example, the watermarking can be extract by doing string comparisons if document W is Window’s word format or html format. The watermarking can be extracted by doing texture comparisons if document is Latex dvi format.

The watermark should have a high bit rate. The bit rate is the bit amount the watermark has. Text watermarking has high bit rate when the text document is a long document; otherwise the bit rate is low. As illustrated by the Section 3, the bit rate of the example is n, and n is the amount of line in text document. The n means the bit amount of a watermark. It is total $2^n - 2$ difference watermarks can be embedded in the document with n lines. The reason for negative 2 is that [000...00] and [111.....111] doesn’t allow use avoiding the watermark collusion. We can see that for the short text document, such as 4 lines text document (n=4) the bit rate only 4, but if 100 lines document the bit rate jump to 100, and have $2^{100} = 1.2 \times 10^{30}$ difference watermark. For more difficult to discovery the location of watermarking, we can use part of the line amount to hide watermark in a long text document. For example in a 100 line document, we only randomly choose 40 lines, and the bit rate is 40. Overall, the text watermarking is limited by its low bit rate when the document is short. The following table shows the bit rate and what this bit rate can possibly embed:

| Lines of the Documents(n) | Bit rate | The content of the watermark | Range |
|---------------------------|----------|---|-------------------------|
| 4 | 4 | [0000] and [1111] can not used | Number :1---5 |
| 7 | 7 | 0x8f,[00...00] and [11..11] can not used | Number: 1--127 |
| 136 | 136 | A test “Author: ChaoLi Ou” in ASCII | Any 17 character |
| 1280 | 1280 | A mono colour 21 ×23 pixel bmp image  | Any 160 byte bmp image. |

One way to solve this problem is for the system to put extra watermark information in a lookup table and only a lookup table entry is embedded as watermark in the document. This would save a lot of bit rate. For example a 24 line text document can provide $2^{24} - 2 = 1.6 \times 10^7$ entries. The open space method can increase the bit rate by embedded bit in every space rather than one space of each sentence.

The watermarked document W shouldn’t affect the performances of the document O. The open space method will not affect the user, because most users won’t recognise the gap difference between the words. The syntactic method will not affect the user if the watermarked document preserves correct grammar. The semantic method will not affect the user, because the synonym substitution will persevere the same meaning of the document O. In order to guarantee synonym substitution, it keeps the same meaning of document O, human is involved to check the document context. The semantic method cannot apply to a text document requiring precise meaning, such as legal document or contract.

In my example, the watermark is seven bit data, and the protection system can use this watermark as an entry of a lookup table which records copyright information to extend the watermark.

The watermark is preserved under routine handling of the watermarked document. This feature is called “robustness”[4]. It has not affected the watermark when the document is distributed by internet. The white space watermark is preserved when cutting from an HTML document and pasting into a Window’s word document. It is also preserved when copy and paste go from Window’s word to HTML. Most of the dvi viewers don’t allow do cutting and pasting in the dvi text document. Some dvi viewers allow the user cutting from dvi text, and then pasting to Window’s word. This operation doesn’t preserve the white space between the line and the word’s gap. I tested Acrobat Reader by viewing a pdf document, and did a cut and past to Window’s word; Window’s word did not preserve the document at all. The result is same as cutting and pasting pdf to html. Some words even cannot transfer. I tested a pdf document :“**The not so short to introduction Latex 2e**” by, Tobias Oetiker, Hubert Partl, Irene Hyna and Elisabeth Schlegl, Version 3.22 ,10 October,2002. In page 77, a word “difficult” in pdf, when cut and paste to Window’s Word is “**di_cult**” and it is “**di♫cult**” in html. The result is that the open space watermark in LaTeX or pdf document is lost when you cut and paste to html and Window’s Word, but the same paragraph looks by far better in Acrobat reader than in Window’s Word. The syntactic method and semantic methods survived any cut and paste operation. I also showed that the text watermark survives compress operation of text document.

The Latex format is doing best in preserve watermark and copyright protection. The Window’s word and html can be changed directly by the user, because the document source code sends to user. The user browser has a function to any html file’s source code. The Latex format is different. It send a dvi document to the user, user can see the document by the dvi view. The W.dvi document is produced by compiling the document W.tex . The source of the document do not sent to user, so the user is difficult to change the contain of dvi document.

5. Conclusion

Text watermarking is a method for copyright protect. It is a reliable and cheap way in most cases, especially for the normal user. However, the algorithm will not survive the attack of a dedicated computer hacker. One way to improve watermark protection is to keep the cost of text document at a reasonable level to prevent it from being an attack target.

Which method is the better of the three discussed is dependent on the document situation. If the document requires accuracy, such as a legal document, then it cannot use the semantic method. In order to guarantee the original document O has the same meaning of the watermarked document W, the protection system has to have someone to make judgements. The person who converts O to W may be a threat to the protection of the system by leaking the document O or by misunderstanding document O. The

disadvantage of syntactic watermarking is the bit rate is low. The reason is that not all sentences have correct grammar and meaning after delicate reordering the words. The open space method has advantage of high bit rate, easy implementations, and perfectly preserves the meaning of the original document. I argue that the open space method with a secret key and the document format Latex or pdf would be the most suitable watermarking method for text document for copyright protection.

References:

- [1] Christian D. Jensen, “*Fingerprinting Text in Logical Markup Languages*” G.I. Davida and Y. Frankel(Eds.): LSC 2001, LNCS 2200, pp. 433-445,2001. © Springer-Verlag Berlin Heidelberg
- [2] Steven H. Low, Member, IEEE, and Nicholas F. Maxemchuk, Fellow, IEEE, *Performance Comparison of Two Text Marking methods* ,IEEE Journal on Selected areas in communications, vol 16, NO 4 May 1998.
- [3] Christian S. Collberg, Member, IEEE Computer Society, and Clark Thomborson, Senior Member, IEEE, *Watermarking, Tamper-Proofing, and Obfuscation–Tools for Software Protection*, IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 28, NO. 8, AUGUST 2002 .
- [4] Clark Thomborson, private communication 20 April 2003
- [5] Mikhail J Atallah, Victor Raskin, Christian F. Hempelmann, Mercan Karahan, Radu Sion, Umut Topkara, Katrina E. Triezenberg. “*Natural Language Watermarking and Tamperproofing*” CERIAS TR 2002-38, F.A.P. Petitcolas(Ed.): IH 2002, LNCS 2578,pp. 196-212,3003. © Springer-Verlag Berlin Heidelberg 2003.
- [6] Saraju P. Mohanty, Dept of Comp Sc and Eng. University of South Florida Tampa, FL 33620, *Digital Watermarking: A Tutorial Review*, smohanty@csee.usf.edu