Electronic Commerce Research, 2: 169–187 (2002)

Watermarking of Electronic Text Documents

MOHAN S. KANKANHALLI and K.F. HAU mohan@comp.nus.edu.sg
School of Computing, National University of Singapore, Singapore 119260

Abstract

With the tremendous development of the Internet, it has become desirable to distribute text documents electronically. However, commercial publishers may be reluctant to offer valuable digital documents online for the fear that they will be re-transmitted or copied illegally. To address this problem, we propose a robust watermarking technique, whereby electronic text documents are fingerprinted with one or more semantics-preserving modifications to the document text. The text modifications may be selected so that multiple copies of the same master document will all have the same meaning. By examining text modifications in an unauthorized copy, one can identify the authorized source and the recepient. In this paper, we present a new method that is accurate, robust against attacks (e.g., the cyber pirate may post only a section or a paragraph of a registered text online), scalable (e.g., a few pages of text to hundreds of pages) and secure (e.g., remove or modify embedded watermark with or without knowledge of watermarking method). This approach could therefore facilitate e-commerce of newspapers, journals, magazines, and in general any electronic text document possessing commercial value.

Keywords: watermark, electronic commerce, copyright protection, electronic text

1. Introduction

The enormous popularity of the Internet and other electronic networks in the 1990s demonstrated the commercial potential of offering text documents through the digital networks. The electronic distribution of information is faster, less expensive, and requires less effort than making paper copies and transporting them. Other factors that favor electronic-information distribution include the ability to use a computer to search for specific information, and the ability to easily customize what is being distributed to the recipients. Also, recipients may be able to choose the desired method of presentation. Consequently, electronic newspapers, magazines and journals are poised to supplement, and perhaps eventually overtake, the current paper distribution networks.

However, the very properties that make "the net" attractive as a distribution medium—ease of manipulating information in an electronic form, also appear to make protection of intellectual property difficult. Once the publisher delivers the digital content upon payment, the customer may offer this content on his web site, post it to a Usenet newsgroup or e-mail it to friends for a lower price or even perhaps for free. Generally, it is easier for a person who obtains an electronic document to forward it to a large group than it is for a person who receives a paper copy of the same document. In addition, electronic copies are more akin to the original than paper copies. When an electronic copy is made, the original owner and the recipient have identical entities. A person with a photocopy of a journal and

a person with the original bound journal may have the same information, but it looks and feels different. Illicit copies of electronic documents are likely to result in major loss of revenues.

Now, digital content publishers are at a crossroad. On one hand, they would very much like to use the web to make higher profits due to a larger customer base, a lower distribution and delivery costs using the web. On the other hand, they will lose revenues due to cyber piracy. Some online publishers, when faced with this dilemma, decide to offer a few pages of popular novels as teasers on the web, meanwhile continue to sell and ship the physical copies to the reader. By doing this, the publishers still make use of the web to reach out to a larger customer base, but the benefits of a lower manufacturing and delivery costs are lost, since the publisher still has to print, bind and deliver the physical book.

The primary goal of information protection is to permit proprietors of digital information (i.e., the artists, writers, distributors, packagers, market researchers, etc.) to have the same type and degree of control present in the "paper world". Because digital information is intangible and easily duplicated, those rights are difficult to enforce with conventional information processing technology. Since commercial interests seek to use the digital networks to offer digital documents for profit, they have a strong interest in protecting their ownership rights. We believe that electronic publishers will only offer valuable documents online when effective techniques to combat cyber piracy are available. In our paper, we will focus on techniques to protect the intellectual properties of online content publishers.

2. Literature review

Protecting intellectual property has received a lot of attention lately, both in terms of revised intellectual property laws [Goldstein, 14], as well as new technology-based solutions. It must be absolutely clear that any technological solution must be backed up with the appropriate legal framework to resolve disputes. We concentrate on the technical aspect of the problem in this paper. Most existing techniques that have been invented to solve the problem of piracy fall into two categories: copy prevention and copy detection.

Copy prevention

Copy prevention schemes include physical isolation of the information (e.g., by placing it on a stand-alone CD-ROM system), use of special-purpose hardware for authorization (e.g., secure printers with cryptographically secure communication paths) [Osawa, 25; Popek and Kline, 27; Shivakumar, 30], and active documents [Kohl, Lotspiech and Kaplan, 19; Popek and Kline, 27; Shivakumar, 30; Sibert, Bernstein and van Wie, 33] (e.g., documents encapsulated by programs) where users may interact with documents only through a special program.

However, it is felt that such protection schemes are inflexible (require special hardware or software), constraining (restricted access to documents), and not always completely foolproof (an OCR program may be used to reformat the document, software emulators can be used to break the prevention schemes).

Copy detection

In view of the shortcomings of the prevention approach, researchers have been concentrating on detecting illegal copies, that is, we make an assumption that most users are honest users and we allow them to access the documents, meanwhile, we concentrate on detecting those that violate the rules. Here when we refer to "copy detection" we include detection of both full and partial copies. There are two approaches in the detection schemes, watermark based and registration based.

Registration-based schemes

In registration-based copy detection schemes, a copy detection server is established [Anderson, 3; Parker and Hamblen, 26]. When an author creates a new document, he registers it at the server. The server could be used as a repository for a copyright recording and registration system [Kahn, 17]. As the documents are registered, they are broken down into small units. Subsequent documents that are produced are compared against the pre-registered documents for partial or complete overlap.

There are a number of ways of detecting duplication of registered documents. We will discuss two methods in detail.

2.1. COPS

In COPS [Brin, Davis and Garcia-Molina, 10], registered documents are broken up into sentences or sequences of sentences, and are stored in the registration server. When a document is to be checked, it is also broken into sentences. For each sentence, a hash table is probed to see if that particular sentence has been seen before. If the document and a previous registered document share more than some threshold number of sentences, then a violation is flagged. The threshold can be set depending on the desired checks, smaller if one is looking for copied paragraphs, larger if one wants to check if documents share large portions. A human would then have to examine both documents to see was it was truly an act of plagiarism.

2.2. SCAM

In SCAM [Shivakumar and Garcia-Molina, 31; Shivakumar, 30], word chunking is used instead of sentence chunking. An index of the word chunks in the vocabulary is constructed and maintained at registration time. Subsequent query documents are broken up into word level as well. A scheme that combines the relative frequency of words as indicators of similar word usage and the cosine similarity rule is used to compare documents. A vector that gives the frequency with which each possible word occurs in the new documents is computed and then compared against "similar" vectors in the repository of registered documents. Threshold limits similar to those used in COPS can be used to test for full plagiarism, subsets or related works.

In comparing the above two schemes, empirical results have shown that word chunking used in SCAM performs better than sentence chunking used in COPS. Word chunking leads to more locality during comparisons, and has the potential to detect finer (e.g., partial sentence) overlap, which may be particularly important with informal documents that may not have a clear sentence structure. However, results also show that SCAM reports more false positives than COPS, where false positives are pairs of documents that are reported to be possible instances of plagiarism, but are not. So true plagiarism cases are indeed detected at the cost of lot of many spurious alarms.

Signature-based schemes

In signature-based schemes, a publisher incorporates a unique and subtle digital watermark into a document when it is given to a user so that when an illegal copy is found, the origins of the document can be traced. At present, there are two main approaches to digital watermarking, namely visible and invisible watermarking.

Invisible watermarking: Line shift encoding

The first method [Brassil et al., 7; Low et al., 21; Low, Maxemchuk and Lapone, 22; Maxemchuk, 23], called line shift encoding, encodes the document uniquely by vertically shifting the locations of text-lines. One bit is transmitted in each line that is moved. During decoding, the digital image of the document is obtained and text-lines are located using horizontal projection profile. The distance between adjacent text-lines is measured. This can be done by either measuring the distance between the baselines of adjacent lines or the difference between centroids of adjacent lines. This method works for formatted documents only.

Invisible watermarking: Word shift encoding

The second method [Brassil et al., 7, 8; Low et al., 21; Maxemchuk, 23], called word-shift encoding, is a method of altering a document by horizontally shifting the locations of words within text-lines to encode the document uniquely. Unencoded lines are included to detect and compensate for nonlinearities that occur in printing and copying. However, this method is only applicable to documents with variable spacing between adjacent words, and because of the variable spacing requirement, decoding requires the knowledge of the space between words in the unaltered document. Again, this is useful only for formatted text documents.

Invisible watermarking: Feature encoding

The third method [Amano and Misaki, 2; Brassil et al., 7; Brin, Davis and Garcia-Molina, 10] is called feature encoding. The formatted document is examined for chosen features, and these features are altered, or not altered, depending on the codeword. Decoding requires a specification of the change in pixels at a feature.

Invisible watermarking: Inter-character space encoding

In inter-character space encoding technique [Chotikakamthorn, 11], inter-character spacing is varied to encode a binary watermark. This encoding scheme is designed to cater for text documents with languages such as Thai that have few or no large inter-word spaces. Not only does this method have wider applicability in terms of supported languages, it also provides more embedding capability than the methods in [Brassil et al., 7; Low et al., 21; Low, Maxemchuk and Lapone, 22; Maxemchuk, 23]). But again, it works for formatted documents only.

Invisible watermarking: High-resolution watermarking

A document file is programmed such that it has two or more components [Adams, 1], with one of the component a watermark object. The watermark object is a high-resolution pattern with some binary symbols encoded inside. The resolution is high enough such that the binary symbols cannot be percieved by the human eye. A processor may be programmed to recognize the pattern, decode the pattern into binary data, and decode the binary data to characters directly interpretable by a user. Information relating to creation and control of a document, signature or the like, may all be encoded independent from the principal image (e.g., text, graphics), to be virtually undetectable by human eyes, yet non-removable by copying methods, including photocopying, scanning, etc.

Invisible watermarking: Selective modifications of the document text

In this scheme [Nielson, 24], electronic documents are fingerprinted with one or more modifications to the document text. The text modifications may be selected so that multiple copies of the same master document will all have the same meaning.

Preparation of a master document for electronic distribution may begin with identification of places in the document for which two or more alternative strings would provide the same meaning. As many such instances would be identified as there are digits in the binary customer number used to identify authorized recipients.

By examining text modifications in an unauthorized copy, one can identify the authorized source. Our work in this paper belongs to this class of techniques. Our method differs in the type of modifications made to the text. We primarily concentrate on changing the punctuation and abbreviations.

Invisible watermarking: Centroid decoding method

To derive the maximum likelihood detector for the centroid method, [Brassil et al., 7, 8; Low et al., 21; Maxemchuk, Low, Maxemchuk and Lapone, 22; Low and Maxemchuk, 20] [Brassil et al., 7; Low et al., 21; Low, Maxemchuk and Lapone, 22; Low and Maxemchuk, 20], the effects of additive profile noise is characterized by the centroid positions. Involved mathematical calculations show that the exact centroid noise

density has a very complicated formula. As it is too complicated to be used for centroid detection, an approximation is proposed. The use of an approximation is justified by sketching its error bounds and real document profiles are used to demonstrate its accuracy. The major conclusion here is that, for typical profiles, the centroid noise is, approximately, zero-mean Gaussian whose variance is easily computable from the original unmarked profile. This approximation helps derive the maximum likelihood detector for the centroid method and its error probability.

Invisible watermarking: Correlation decoding method

The correlation method [Low et al., 21; Low and Maxemchuk, 20], treats a profile as a waveform and decides whether it originated from a waveform whose middle block has been shifted left or right. Both the maximum likelihood detector and its error probability are computed. Detailed information about this method can be found in [Van Tress, 34].

Visible watermarking

Meanwhile, visible watermarking has also been investigated by some research groups, though the target of watermark has never been text documents. The watermarks are mainly used to provide copyright protection for high quality images and video.

The IBM group has done some research on visible watermarking [Braudaway, Magerlein and Mintzer, 9]. A visually unobstrusive watermark is embedded into a large area of the image by modifying pixel luminance. A pixel in the image is darkened or brightened according to where the brightness of the corresponding mask pixel is located in the linear brightness scale. Randomization is added to the strength and the location of the watermark to make it less vulnerable to automated removal.

Another visible watermarking technique is proposed in [Kankanhalli, Rajmohan and Ramakrishnan, 18], whereby the intensity of the watermark is varied in different regions of the image depending on the underlying content of the image. It is achieved by first analyzing the image content such as textures, edges, and luminance before providing a just noticeable distortion level for each block. Using this measure for varying the strength of the overlaid secondary image ensures that it is perceptually uniform over different regions of the image and the process can be automated over a wide range of images.

The authors have done some research using visible watermarking approach adopted from the IBM group [Braudaway, Magerlein and Mintzer, 9], to be used on text documents [Hau, 16]. A text document is converted to an image file first, and a visible watermark is inserted by manipulating the pixel luminance values. A simple histogram equalization is performed on the watermark to the noise and hence a better-quality watermarked text image is produced. Randomization is added to the strength and the location of the watermark to make it less vulnerable to automated removal.

3. Marking an electronic text document with a robust watermark

It is easy to build a system that can retrieve the embedded watermark in an illegal document that is not tampered with. The full watermark can be retrieved and by comparing the

watermark detected against the watermarks registered in the database, we can easily find the rightful owner of the detected watermark.

However, in the real world, this is not the case. The cyber pirates usually tamper with the document such that the resultant document is not an exact copy, but a similar copy to the original document. For instance, the cyber thief may copy only a section, chapter or paragraph out of an entire document. He may also modify the document by inserting (or deleting) sentences and words not in the original document. To complicate matters, instead of adding sentences randomly, he may get hold of several legal copies of the same document by colluding with the legal owners of the document and produce another copy of the document by using different portions of the document from the various copies available. The last form of attack we can contrive is that a cyber pirate may possess a copy of a document legally at first. This implies the copy he is holding on has already been watermarked. Subsequently, he may watermark the copy again before posting it on the web or distributing it illegally, since the original watermark has been overwritten and there is no way to trace the one who leaked out the document initially.

The process of watermarking any document begins with the entry of the rightful owner of the to-be watermarked document into the database. The desired ASCII string of characters which serves as the watermark to be embedded into the document is converted into a string of binary digits, for example "0011010101". Subsequently, the document is broken down into sentence levels and each sentence is passed into a punctuation module followed by a word module. The punctuation module will detect suitable sentences for which two or more alternative strings would provide the same semantic meaning. Every suitable sentence will have one bit of code encoded. During detection, each sentence will be broken up into tokens. Detection is done mainly by recognizing the key words and some regular syntactic structures in the English language as follows:

- if Token.equals (any Coordinate Conjunctions);
- if Token.equals (any Relative Pronoun);
- if Token.equals (any Conjunctive Adverb with only one word);
- if Tokens.equals (any Conjunctive Adverb with more than one word);
- if Token.equals (commas) AND previousToken.equals (inverted commas or colon).

The sequence of detection will be done according to the above order. Detection stops when the first identification of keyword/s is found. This will be followed by the appropriate modifications to the sentence, according to the set of punctuation heuristics that we have drawn out in this sequence:

- Place a comma before a coordinate conjunction linking two independent clauses to represent 1 and omit it to represent 0. Leave the sentence untouched if there are any coordinate conjunctions found before the word count reaches five or if the coordinate conjunction is used as the last word in the sentence.
- Place a comma after a conjunctive adverb linking two sentences to represent 0 and omit it to represent 0. Leave the sentence untouched if the adverb is the last word in a sentence.
- Place a comma before a relative pronoun linking two clauses to represent 1 and omit it to represent 0. Leave the sentence unchanged if the relative pronoun appears as the first word of the sentence.

- Replace the colon use in direct quotes with comma to represent one, do the reverse to represent 0.

After a sentence goes through the punctuation analysis, it is passed into the word module. The word module will detect sentences suitable for which two or more alternative strings would provide the same semantic meaning. Every suitable sentence will have one bit of code encoded. Detection is done mainly by recognizing the key words that are already stored in the word database, and once a key word or phrase is found in the sentence which also exists in the database, detection stops. This will be followed by the appropriate modifications to the sentence, according to the set of word heuristics that follow:

- Substitute the abbreviations with their original meanings to represent 1 and the reverse to represent 0.
- Substitute "must" with "should" to represent 1 and the reverse to represent 0.
- Substitute word phrases with more concise words to represent 1 and the reverse to represent 0.
- Substitute abbreviation "a.m." with "A.M." to represent 1 and the reverse to represent 0. Likewise for "p.m.".

Hence, we can see that any sentence can be watermarked with 1 binary code or 2 binary codes, or it can be untouched if it is seemed unsuitable for any syntactic or semantic changes.

After the whole text document is encoded completely with the binary watermark, the watermark is stored against the certified owner of the document in the database. The encoded string could possibly identify the recipient also, if desired.

4. Detecting a watermark in an electronic text document

Detection of watermark is done together with the detection of all similar sentences present in any registered document found in the database.

To start, the first registered text document is broken down into sentence levels and each sentence is stored into a table indexed by the order in which the sentence appears. The same treatment is also done for the targeted text document in which the watermark is to be detected. The two indexed tables are then passed into a detection module, whereby each sentence in the registered text in which the current table pointer is pointing is compared with the sentence in the targeted text. A similarity flag is raised, if the number of identical words found between the two sentences that are compared are more than half of the total number of words in the sentence that belongs to the registered text. If the similarity flag is raised, the particular sentence in the targeted text is first decoded for syntactically embedded codes based on the following heuristics in order:

- If a comma before a coordinate conjunction linking two independent clauses is found, the embedded code is returned as 1. If no comma is found, return the embedded code as 0. No embedded code will be returned if there are any coordinate conjunctions found before the word count reaches five or if the coordinate conjunction is used as the last word in the sentence.

- If a comma after a conjunctive adverb linking two sentences is found, the embedded code is returned as 1. If no comma is found, return the embedded code as 0. No embedded code will be returned if the adverb is the last word in a sentence.
- If a comma before a relative pronoun linking two clauses is found, the embedded code is returned as 1. If no comma is found, return the embedded code as 0. No embedded code will be returned if the relative pronoun appears as the first word of the sentence.
- If a comma is found before the start of a direct quote, the embedded code is returned as 1. If a colon is found instead, return the embedded code as 0.

Decoding for syntactic heuristics stops when either a "1" or a "0" is returned. If no code is returned after all four heuristics are tested, a "−1" is returned, showing that a similar sentence is found, but no embedded code is detected.

Subsequently, the sentence is decoded for semantically embedded codes based on the following heuristics in order:

- If the original meanings of the abbreviations recognized by the word database are found in the sentence, returned the embedded code as 1, else if recognized abbreviations are found, return the embedded code as 0.
- If the word "should" is found, return the embedded code as 1. If the word "must" is found, return the embedded code as 0.
- If shorter and more concise words to long word phrases recognized by the word database are found, return the embedded code as 1, else if the longer word phrases are found, return 0.
- If special abbreviation "A.M." is found, return the embedded code as 1, else if "a.m." is found, return 0. Likewise for "p.m.".

Decoding for semantic heuristics stops when either a "1" or a "0" is returned. If no code is returned after all four heuristics are tested, a "−1" is returned, showing that a similar sentence is found, but no embedded code is detected.

This procedure repeats until every sentence in the registered text is tested, and all the returned embedded codes except "−1" are appended to form a detected watermark string. Each symbol "|" is used to represent every sentence found in the registered text which embeds a binary code but is not found in the targeted text. Hence, every missing sentence can yield at most 2 "|" in the watermark string retrieved. To summarize, the eventual detected watermark string will be the watermark retrieved when compared to a particular registered text document only. Hence, it can only be used for analysis with respect to that particular document.

After the watermark string is retrieved, the whole procedure is repeated with the next registered document in the database, and a new watermark string will be obtained.

5. Enhancing the robustness of the watermark

To make our watermarking scheme more robust, we have incorporated some additional features. First of all, every ASCII character that is converted into a binary string is appended with a string of 8 "1"s as trailer. This is done to insert an easily recognizable

artificial pattern into the to-be-embedded watermark. By doing so, if we are given a document to be watermarked, we can do a preliminary check on the existing code pattern on that document, and if the strings of "1" are found to exceed our threshold which is set arbitrarily, we can raise an exception and conclude that we have reasonable confidence that the document given is already watermarked and any further attempts to watermark will be regarded as tampering. Furthermore, to instill more randomness and unpredictability, a random number is generated and converted into binary digits and appended to the earlier binary string.

Next, we append the to-be-embedded watermark string with 32 bits of cyclic redundancy code to form the final watermark string. During detection, the string of codes retrieved is divided with some predetermined binary number. If the remainder obtained is a string of "0"s, we can be assured that there is no tampering previously in the targeted document, else, we can immediately declare that some tampering has been done to the document even before we do any further analysis.

Lastly, some design effort has been geared towards collusion detection. After we retrieve the binary watermark string, we analyze each position of the string and compare it at the same position with one of the registered watermark in the database. If for some positions, the binary codes found in the registered watermark differ from the retrieved watermark in the given document, these "positions of difference" will be compared with other registered watermarks. If we can find one or more registered watermark/watermarks which have similar codes in these "positions of difference", we can raise a flag indicating potential possibility of collusion.

To do a cyclical redundancy check (CRC), the string of binary digits is retrieved from the text and is divided by the same pre-defined CRC polynomial again. If the remainder is not zero, then there is a suspected tampering of the embedded watermark.

6. Testing our technique with simulated cases

6.1. Experiments

Evaluating the quality of similarity measures is tricky, since these measures need to approximate a human's decision process. Also, even two humans may not agree on whether two documents are similar. Ideally, if a benchmark database of copyright violations were available, we could evaluate our similarity measures against these data. Unfortunately, no such benchmark exists. Thus, for our experiments, we start with watermarking a set of five documents. We then modified these documents such that that they have "substantial" overlap. We also modified 2 un-watermarked documents such that they consist of portions of watermarked segments from documents of similar content. We then fed these documents into our system, and the system determined their authenticity based on apparent similarities. At the same time its original owner is determined by retrieving the watermark embedded.

Human classification

For our experiments, we first classify documents based on the following predicates. We shall stress that this classification was very subjective, but was our best attempt to identify documents that most humans would agree are closely related.

1. *Partial resemblance*. If a piece of document includes some clearly recognizable portions of another document, the document pair will satisfy the predicate.
2. *High resemblance.* This is similar to the previous category, except that a document includes large portions of another document, possibly interspersed with other text.

Data sets

For our experiments, we use 2 datasets of documents. The first dataset consists of five documents. For each document, we fabricated three documents based on the original document, with two exact watermarked duplicates of it, and one highly resembling the original, according to our human classification.

The second dataset consists of two documents. In making a document of partial resemblance, we look for a document in dataset 1 that has similar or related content to our document in dataset 2. We then insert sections or sentences from a watermarked copy of it into this document in dataset 2. We thus create two documents that have partial-resemblance.

Fabrications of the original documents

For each of the five pieces of documents in dataset 1 used in our experiment, we produced two watermarked copies using our prototype system. Either one of them can be used to be tested for exact replicate to the original copy. To create a new document that has high resemblance, we removed certain sections of the documents or alternate sentences of a paragraph after reading the document manually. Furthermore, we reordered some paragraphs of the documents to reduce the visual resemblance of this new conjured piece of document to the original one. Whenever possible, we inserted new sentences into the document to reduce apparent similarity. We perform these actions only when they do not affect the overall coherency of the documents. We emphasize again that our decisions were subjective and they would differ from individual to individual. In an attempt to simulate an illegal collaboration to defeat our watermarking system, the various sections of a new document is made up by lifting the respective sections from either of the two watermarked documents in a random manner. Partial resemblance documents are produced in a similar manner, except the degree of changes and manipulations are more drastic and the similar portions only contribute to a minor part.

6.2. Results

Our experiments were carried out on a 233 MHz Pentium MMX laptop with 32 MB RAM.

Exact replicates

When we performed decoding on an exact replicate of the various documents, we obtained a very satisfactory result. For Document 1, we found 148 similar sentences in the replicate, which is the total number of sentences in the original document. Also, out of the 79 binary codes encoded in the document, we managed to retrieve the full 79 binary codes without error. For Document 2, we found 273 similar sentences out of a possible of 274, and 140 binary codes out of a possible of 141. For Document 3, we found 261 similar sentences out of a total of 262, and retrieved the full 209 binary codes embedded. For Document 4, we found 351 matching sentences out of a total of 354, and retrieved 99 out of 100 binary codes. Finally, in Document 5, we have a 197/197 sentence matching, and retrieved the full 72 binary codes embedded.

However, in the replicate of Document 5, an alert flag on illegal watermark manipulation is raised. This means that when our system performed cyclic redundancy check on the binary watermark strings recovered, an anomaly was detected. The implication of this is that there is unintended modification to the original watermarks embedded in the replicate, and we may not be able to identify the original owner of the replicate with full confidence. Instead, a range of possibilities is returned by the system, and the owner/owners who have the highest proportion of similarity when comparing their assigned watermarks with the recovered one will be classified as the most probable culprits.

Hence, we can conclude that our prototype system works extremely well in recognizing ownerships of untampered replicates, where the full string of binary watermark can be recovered.

High resemblance

With respect to the modifications that we have made to conjure new documents that "look" similar, our prototype system is able to successfully obtained figures that can prove their identity with credible probabilities. For the high resemblance replicate of Document 1, 116 out of 148 sentence matches were found, and 62 binary codes were retrieved. In Document 2, 261 similar sentences were found and 130 binary codes were retrieved. Document 3 had 236 similar sentences in its highly-overlapped replicate, together with 188 binary codes retrieved. For Document 4, 317 similar sentences were found together with 88 binary codes. Finally, Document 5 has a 160/197 sentence-matching ratio, with 65/72 binary codes retrieved.

Our results and figures have shown that our system is able to obtain enough "clues" in terms of similar sentences found and the partial watermarks retrieved to provide a list of registered owners suspected of illegal plagiarism and/or collusion. Our algorithm and heuristics used works in such a way that all possible combinations of owners were tested against the clues retrieved, and the combinations will be reported if a match is found. The nature of our search algorithm dictates that combinations involving less owners have a higher probability of being the actual culprits than combinations with more owners in terms of collusion. With a document having high resemblance, it is easy to sieve out the most probable culprit given the extensive information recovered.

Partial resemblance

We did decoding of partial resemblances in two of our documents, Document 1 and Document 2. The regions of similarity are obvious to the human judgment, although the total proportion of similarity is not significant. However, it is worthwile to take note that our system is able to catch hold of most cases of similarity that is obvious to the human being. For Document 1, we found 27 similar sentences and retrieved 11 binary codes in the newly conjured copy and for Document 2, we have 32 matches and 20 binary codes are retrieved. When we try to watermark both conjured documents, the preliminary tests in our system could not tell that some part of the documents have already been watermarked and thus should not be able to be watermarked again.

Due to the limited information recovered, our system is unable to provide a definitive answer to who the legal owner that inserted parts of his given document to create a new one is. Instead, after the document undergoes a series of checks, all the possible owners, together with their potential collusion partners, are shown, each with almost equal probability of being the correct one. If the number of authorized recipients is huge, for example, in the hundreds, we may receive a report from our system indicating that all the authorized recipients, and all the possible combinations of these recipients are in the suspect list, each with an almost identical probability of being the one. Thus, the identity of the culprits cannot be established on the basis of this information.

The experimental results are compiled and shown in Table 1.

Encoding time

In our experiments, we have performed encoding of watermarks into the respective documents 5 times. Using the system clock, we have compiled the relationship between time

Table 1.

	OVERLAP	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
Percentage of	Some	18.3232%	9.1636%			
similar sentences found	High	78.37837%	95.2727%	90.4580%	89.5480%	81.2182%
found	Exact	100.0%	99.63636%	99.61832%	99.15254%	100.0%
Percentage of	Some	13.9240%	11.542%			
binary watermark	High	78.4810%	92.9078%	89.952%	88.0%	90.278%
retrieved	Exact	100.0%	99.2907%	100.0%	99.0%	100.0%
Can the system recognize	Some	No	No			
the artificial pattern in the	High	Yes	Yes	Yes	Yes	Yes
string of binaries retrieved?	Exact	Yes	Yes	Yes	Yes	Yes
Is there any	Some					
tampering to	High					
the watermarks?	Exact	No	No	No	No	Yes
Has the correct	Some	Inconclusive	Inconclusive			
owner of plagarism	High	Yes	Yes	Yes	Yes	Yes
identified?	Exact	Yes	Yes	Yes	Yes	Yes

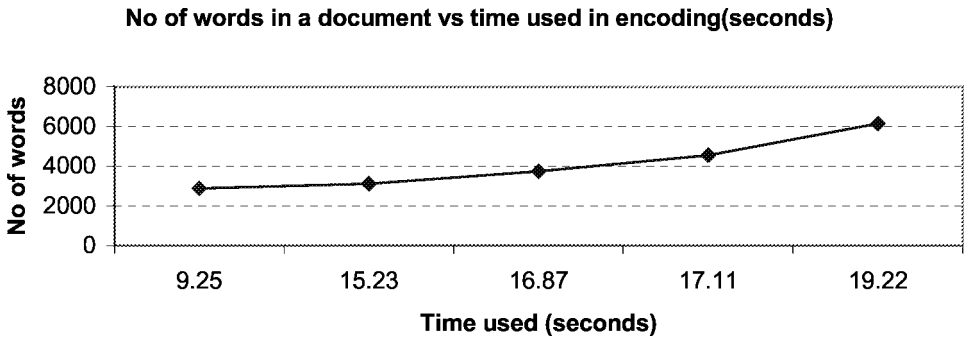

Figure 1.

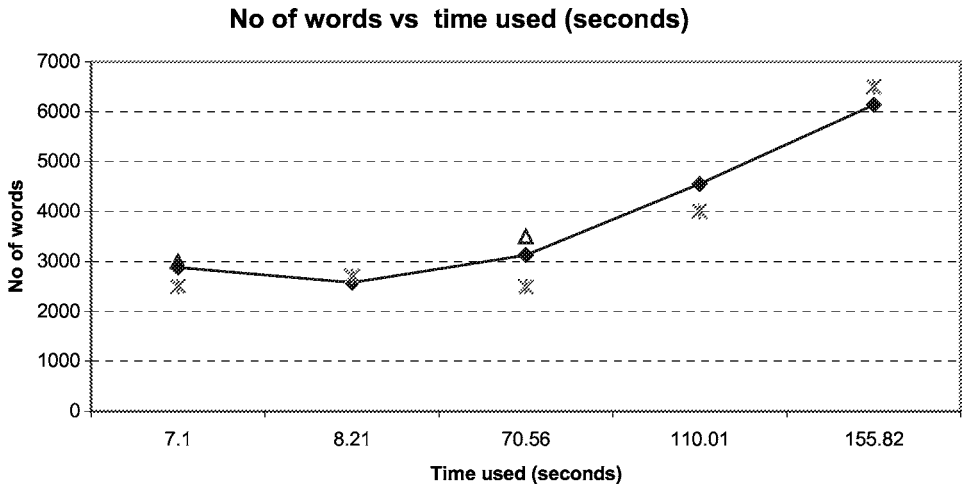

Figure 2.

taken to encode a document and the length of the document into the graph shown in Figure 1. Intuitively, we would expect a direct relationship, since a longer document means that more traversals of the document and most likely more encoding to be done.

Decoding time

We have performed decoding of 12 documents in our experiments altogether. Similarly, we have make use of the system clock to help us keep track of the relationship between time taken to decode a document and the length of the document and compile our results into the graph shown in Figure 2. Intuitively we would expect a direct relationship as well, since a longer document will mean that most probably more binary codes encoded and hence more CPU cycles need to be used to retrieve the codes. However, we notice a slight dip in the graph. We will attribute the cause of the dip to the sentence structure of

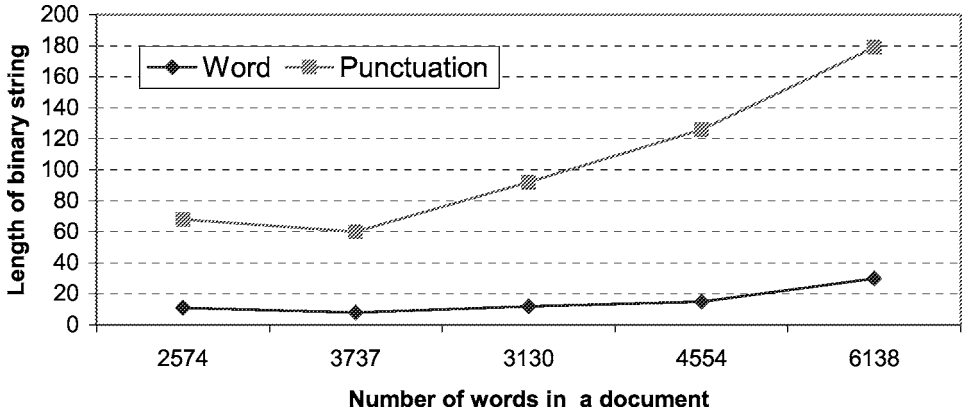

Figure 3.

the documents involved. The documents that have a decoding time of about 7.1 seconds are in fact newspaper articles from the Asian Wall-street Journal. Newspaper articles tend to have shorter sentences compared to technical papers. Hence, the number of iterations performed on each sentence is smaller. This contributes to the overall shorter time spent on decoding, even when the length of the document is longer.

Word heuristics vs punctuation heuristics

To have some insight into how frequently word heuristics and punctuation heuristics are used in encoding of watermarks, we added in an extra utility program that returns the breakdown of the number of times word heuristics and punctuation heuristics is used to embed a binary code in a document. The result is shown in the graph in Figure 3. We can see that given a piece of document, the number of times punctuation heuristics is used, denoted by the length of the binary string that is encoded using punctuation heuristics, is significantly longer than that of word heuristics, and the difference gets wider as the document becomes longer. This is an important observation, since we should now try to expand our word heuristics to remove the gap before we face situations when there are limited syntactic modifications possible, such as lecture notes or newspaper articles, where sentences are either short or in pure clause form.

7. Testing with two cases of real-life plagiarism

Real life plagiarism

Our two cases of plagiarism are both drawn from undergraduate students copying material from multiple web-sites with almost no modifications to the original sentence structures. After minimal reorganization, these materials were published in the web as student reports to meet some specific course requirement. No relevant acknowledgments were given to the source web-sites and these students took full credit for the work.

Case study 1 Using a plagiarism detection engine at `www.findsame.com`, we looked for known cases of plagiarism. The first case of plagiarism involves an undergraduate student who needs to write a report on the Ariane 5 (a satellite) explosion. The name of the student and the university have been kept anonymous.

The report is found to have be vastly similar to another piece of article published on the website `www.math.rpi.edu/~holmes/Numcomp/Misc/siam.ariane.html`. In fact, the detection engine reported a similarity of 73%.

To start, we got hold of the original piece of document from the website mentioned above, detected its default watermark and registered it in our database. Then, we ran our prototype system with the suspected report.

According to our system, 85 similar sentences are found between the 2 documents. This translates to a 90% similarity. Furthermore, the watermark retrieved from the suspected student report has 64 identical binary codes out of a possible of 71when compared to the watermark registered in the database. Here, we define two binary codes to be identical if they have the same binary value and they are at the same position in their respective binary string. These two pieces of evidence are strong enough to convince us that it is not mere coincidence that we have two instances of highly similar articles, since the probability of a pure coincidence is rather low.

Case study II The second case of plagiarism was obtained in a similar manner as the first. Using the same search engine, we located a case whereby a student lifted material from multiple web-sites and did a report on Internet firewalls by purely reorganizing these materials. A crawl by the detection engine yielded a dozen URLs in which there is recognized similarity. On average, about 10 sentences were lifted from each web-site. We selected the articles at three of these web-sites at random, detected their default watermarks and registered them into our database. Subsequently, we ran our system with the suspected report.

On comparing the student's report with the first original article which was published at `http://core.dynip.com/books/FireWalls\%2520Complete/chap07.htm`, our system found 12 similar sentences out of a total 469. This is about 2.56% similarity. The sentence numbers of all the similar sentences found were displayed. This made it very easy for us to compare the origins of the similar sentences found. Meanwhile, a very short portion of the watermark embedded (10 binary codes out of 410) was retrieved, and when we compared these 10 binary codes with the registered watermark in our database, we find 9 identical codes. Due to the limited evidence gathered in a single comparison and the presence of a mismatch in one of the binary code retrieved, the system decided to flag a negative result.

The second comparison was done between the student's report and an original article retrieved from the site `http://www.nttc.org/doc/firewall/html`. A total of 11 similar sentences were found from a total of 95 sentences (11.58% similarity). 11 binary codes from a possible of 102 were retrieved, and all of them were found to be identical to the watermark registered in our database. Here, in an absence of any mismatch, the system positively flagged a suspected case of plagiarism.

The third comparison was done with the article retrieved from the site `http://itmweb.com/essay534.htm`. A total of 7 sentences were found similar out of 294. This is equivalent to about 2.38% similarity. 8 binary codes were retrieved and all were identical to the registered watermark in the database. Again, in the absence of mismatches, the system returned a positive conjecture.

If we manually examine the sentence numbers of all the similar sentences returned in the above 3 comparisons, we find that there is little overlapping of these sentence numbers. This reinforce our belief that the student had lifted different material from different web-sites and combined them to produce a report with minimal rephrasing. The partial watermarks retrieved served as supporting evidence against the argument that all these similarities happened as a result of pure coincidence.

8. Conclusions and discussion

The proliferation of the Internet access devices in the last few years has made it very desirable for digital content publishers to make Internet their primary medium of document distribution. However, the high probability of illegal transmission or copying of the online documents has been the major obstacle to the wide acceptability of electronic publishing and dissemination. In this paper, we have presented a new text watermarking technique which is accurate, robust against attacks, scalable and secure. The core idea of the technique is to use semantics-preserving syntactic transformations of a text document in order to embed a watermark.

There are several directions for future work. First, we note that the time taken to decode a piece of document takes significantly longer than the time taken to encode the same piece of document. This is because during decoding, every sentence in the document has to be compared with each sentence of all registered documents in the database. This is done to detect similarity of sentences regardless of the order of appearance of these sentences. For example, when comparing a 200-sentence long document with a 300-sentence long document, the total number of comparisons is 60000. To improve on this, we may need to think of more efficient ways of comparison, such as hashing, to reduce the number of iterations and thus the computation time needed to produce a similarity table.

Our prototype system detects similarity by comparing the length of two sentences and the number of common words that appear in these two sentences. Empirical testing shows that this method gives a reasonable amount of accuracy. To improve our system, we are considering developing a more elaborate ways of testing, such as the algorithms used in SCAM [Shivakumar and Garcia-Molina, 31; Shivakumar, 30], so that we may achieve full accuracy eventually.

At present, our system can only accept ascii-coded text files for encoding and decoding. Hence, a reasonable area of improvement will be to extend our system such that it can directly work on Microsoft word files or PDF files as well. This is not difficult to achieve, provided one is able to program and alter the formats of these file extensions.

Finally, we are aware that the effectiveness of our watermarking algorithm depends largely on whether the document has suitable places to insert the watermarks. This is

an area of some uncertainty, since the variability of different text documents does not lead to any heuristic on the number of suitable places available to insert watermark bits per unit number of words. Hence, what one can do is to judge by experience. For example, it is impractical to assume that each sentence in a document will have at least one place suitable to encode a watermark. In view of this, the English language structure needs to be deeply analyzed to look for more heuristics to embed watermarks in between words in a sentence. One caveat is that for certain literary documents, the authors may not be willing to allow alteration of the writing for esthetic reasons. Our watermarking technique cannot be used for such cases.

While there are several avenues for improvement, our method is useful for watermarking general text documents by syntactically changing the text while preserving the semantics. The biggest advantage of our method is that it is robust against the OCR attack which defeats most of the previous work since they are limited to formatted text documents only.

Acknowledgments

We would like to thank the anonymous referees for their very insightful comments which have significantly improved the quality of the paper.

References

[1] Adams, P.M. (1999). "Media Independent Document Security Methods and Apparatus." US patent number 5,974,548, October.

[2] Amano, T. and D. Misaki. (1999). "A Feature Calibration Method for Watermarking of Document Images." In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, September.

[3] Anderson, C. (1991). "Robocops: Stewart and Feder's Mechanized Misconduct Search." *Nature* 350(6318), 454–455, April.

[4] Baird, H.S. (1992). "Document Image Defect Models." In *Structured Document Image Analysis*, Berlin: Springer, pp. 546–556.

[5] Bender, W., D. Gruhl, N. Morimoto, and A. Lu. (1996). "Techniques for Data Hiding." *IBM Systems Journal* 35, 313–336.

[6] Brassil, J.T., S. Low, and N.F. Maxemchuk. (1999). "Copyright Protection for the Electronic Distribution of Text Documents." *Proceedings of the IEEE*, 1181–1196.

[7] Brassil, J., S. Low, N. Maxemchuk, and L. O'Gorman. (1995a). "Electronic Marking and Identification Techniques to Discourage Document Copying." *IEEE Journal on Selected Areas in Communications* 13(8).

[8] Brassil, J., S. Low, N. Maxemchuk, and L. O'Gorman. (1995b). "Hiding Information in Document Images." In *Proceedings of the 1995 Conference on Information Sciences and Systems*, March, pp. 482–489.

[9] Braudaway, G.W., K.A. Magerlein, and F. Mintzer. (1996). "Protecting Publicly-Available Images with a Visible Image Watermark." In *SPIE/IS&T Intl. Symp. on Electronic Imaging Science and Technology, Proc. Optical Security and Counterfeit Deterrence Techniques*, 126–133.

[10] Brin, S., J. Davis, and H. Garcia-Molina. (1995). "Copy Detection Mechanisms for Digital Documents." In *Proceedings of the ACM SIGMOD Annual Conference*, San Francisco, CA, May.

[11] Chotikakamthorn, N. (1998). "Electronic Document Data Hiding Technique Using Inter-Character Space." In *Proceedings of the 1998 IEEE Asia–Pacific Conference on Circuits and Systems*, November 24–27, pp. 419–422.

[12] Choudhury, A.K., N.F. Maxemchuk, S. Paul, and H. Schulzrinne. (1995). "Copyright Protection for Electronic Publishing over Computer Networks." *IEEE Network* 9(3), 12–21.

[13] Garcia-Molina, H., S.P. Ketchpel, and N. Shivakumar. (1998). "Safeguarding and Charging for Information on the Internet." In *Proceedings of International Conference on Data Engineering (ICDE'98)*, Orlando, FL, February.

[14] Goldstein, P. (1996). *Copyright's Highway*. Hill and Wang.

[15] Griswold, G.N. (1993). "A Method for Protecting Copyright on Networks." In *Joint Harvard MIT Workshop on Technology Strategies for Protecting Intellectual Property in the Networked Multimedia Environment*, April.

[16] Hau, K.F. (1999). "Visible Watermarking of Formatted Text Documents." UROP Project Report, School of Computing, National University of Singapore, July.

[17] Kahn, R.E. (1992). "Deposit, Registration and Recordation in an Electronic Copyright Management System." Technical Report, Corporation for National Research Initiatives, Reston, VA, August.

[18] Kankanhalli, M.S., Rajmohan, and K.R. Ramakrishnan. (1999). "Adaptive Visible Watermarking of Images." In *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, June, pp. 568–573.

[19] Kohl, U., J. Lotspiech, and M.A. Kaplan. (1997). "Safeguarding Digital Library Content and Users." *D-Lib Magazine*, September.

[20] Low, S.H. and N.F. Maxemchuk. (1998). "Performance Comparison of Two Text Marking and Detection Methods." *IEEE Journal on Selected Areas in Communication* 16(4), 561–572.

[21] Low, S.H., N.F. Maxemchuk, J.T. Brassil, and L. O'Gorman. (1995). "Document Marking and Identification Using Both Line and Word Shifting." In *Proceedings of Infocom '95*, April.

[22] Low, S.H., N.F. Maxemchuk, and A.M. Lapone. (1998). "Document Identification for Copyright Protection Using Centroid Detection." *IEEE Transactions on Communications* 46(3), 372–381.

[23] Maxemchuk, N.F. (1994). "Electronic Document Distribution." *ATT Technical Journal* 73–80.

[24] Nielson, J. (1999). "Fingerprinting Plain Text Information." US patent number 5,943,415, September.

[25] Osawa, Y. (1999). "Sony Corporation." Available at `http://www.sony.co.jp/soj/CorporateInfo/CHANCENavigator/fo6.html`.

[26] Parker, A. and J.O. Hamblen. (1989). "Computer Algorithms for Plagiarism Detection." *IEEE Transactions on Education* 32(2), 94–99, May.

[27] Popek, G.J. and C.S. Kline. (1979). "Encryption and secure computer networks." *ACM Computing Surveys* 11(4), 331–356, December.

[28] *Proceedings of the IEEE*. (1999). "Special Issue on Watermarking." 1079–1196, July.

[29] "Science and Technology of Data Hiding." Available at `http://www.trl.ibm.co.jp/projects/s7730/Hiding/ethsci_e.htm`.

[30] Shivakumar, N. (1999). "Detecting Digital Copyright Violations on the Internet." Ph.D. Thesis, Department of Computer Science, Stanford University, August.

[31] Shivakumar, N. and H. Garcia-Molina. (1995). "SCAM: A Copy Detection Mechanism for Digital Documents." In *Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries (DL'95)*, Austin, TX, June.

[32] Shivakumar, N. and H. Garcia-Molina. (1996). "Building a Scalable and Accurate Copy Detection Mechanism." In *Proceedings of 1st ACM Conference on Digital Libraries (DL'96)*, Bethesda, MD, March.

[33] Sibert, O., D. Bernstein, and D. van Wie. (1998). "Securing the Content, not the Wire, for Information Commerce." Available at `http://www.intertrust.com/architecture/stc.html`.

[34] Van Tress, H. (1968). *Detection, Estimation, and Modulation Theory*, Vol. I, New York: Wiley.