2009 International Conference on Computer Engineering and Applications
Proc .of CSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

The Design and Application of the Threshold Birthmark Algorithm based on Mutual Information Gain Rate

Li Bin
State Key Lab. of Integrated Service Networks , Xidian University. China

Abstract. In order to resist the damage of digital text works in high BER networks or illegal copy in publication, we propose a robust feature-extracting algorithm based on threshold technology and then construct a new threshold birthmark algorithm. The information gain rate has been used to veraciously measure the importance of the feature in feature-extracting, and threshold(t,n) technology has also been used in constructing birthmark to reduce the cost and achieve good tolerance for different conditions. So the application based on this algorithm can overcome the text defect comes from bad channel or manual malicious modification. The method has been proven IND-CCA2 secure and robust. At the end of this paper we introduce a scheme of copyright protection based on this method.

Keywords: Copyright Protection, Birthmark, Threshold, Feature Extraction.

1. Introduction

Because of the openness of the Internet, the copyright of the digital-text works are faced with many serious threat, so how to protect the copyright of digital-text works has been concerned by more and more researchers. As is known to all, there are usually some superfluous data in the design of the encoding format of multimedia works like sounds and images, therefore, the designers are able to use these superfluous data to store copyright information [3,4], which has relatively less influence on the revivification of sound and image information. However, the text works are realized by the natural language, so any tiny modification will affect its integrity and accuracy. Therefore, the conventional digital watermark technology cannot meet the needs of protecting the copyright of digital text works.

The natural-language based digital text copyright protection technology usually adopts Tamada birthmark technology [1,2], and its principle is to use the statistical methods to select the text, and to make the uniqueness of the digital-text works through a variety of features, which does not have any influence on such works. The core of the birthmark is the attribute-retrieve algorithm which contains many technologies such as TF-IDF, Information gain, Expected Cross Entropy and Mutual Information. Information theory is committed to the nondeterminacy of the information during the transmission. After fifty years information theory is also a fully-fledged theory. In birthmark constructing, information theory and information gain is a good method to measure the importance of the attributes.

Usually the decision system should be more robust to resist the damage of the information. Here exists a dilemma, that is, if improving the dimension of text-feature sub-space, then a great deal of undesired noise will be involved in, which lowers the efficiency of extracting algorithm; if reducing the dimension of text-feature sub-space, then the classification accuracy will also be reduced. Therefore, in addition to choosing the feature retrieving method, how to select the feature sub-space dimension is also the one of the

Li Bin. Tel:86-29-88249710,86-29-13359215979
sliverstone0708@hotmail.com

key issues which determines the practicality of the copyright protection program. Using the idea of threshold signature, we propose a dynamic and adjustable feature-selecting algorithm, and forming a threshold (t,n) copyright protection program of the digital-text works. The program provides users with a flexible way in the choice of text-feature sub-space dimension. Through the dimensional ratio between (t,n) threshold-controlled vector space and the feature sub-space, taking this to realize the digital copyright protection program with different quality-cost ratios, the data-missing problem of text works in network transmission or under malicious modification.

2. Preliminaries on birthmark

2.1 Definition of birthmark

Definition 2.1 (duplication relations) supposes that *Texts* is a collection of natural language texts, with *p* and *q* as the natural language texts. $\cong$ is used to refer to the equivalence relation of *p*, *q* on *Texts*, and it is made to satisfy the following conditions: For the *p*, *q* $\in$ *Texts*, $p \cong q$, and only if *q* is a copy of *p*, then $\cong$ is regarded as a duplication relation. Due to the unreliability of network transmission, there often exist tiny differences between the copy and the original, therefore, *q* can be seen as a semantically approximate product of *p* after being transformed, and it is also considered that *q* and *p* mean basically the same.

Definition 2.2 (birthmark) supposes *p* and *q* are natural language texts, *k* is the secret key, and *f* is the function used to select feature information from the text. If *f* (*p*, *k*) satisfies the following two conditions:

(1) *f* (*p*, *k*) can be selected only from *p* with the help of *k*.

(2). If $p \cong q \Rightarrow f(p,k) = f(q,k)$, then *f* (*p*, *k*) is the birthmark of *p*.

Figuring out the birthmark must have the following two properties:

(1) Confidence: Suppose *p* and *q* are two inter-independent natural language texts, and *k* is the secret key. If *f* (*p*, *k*) ≠ *f* (*q*, *k*), then *f* can be called confident.

(2) Robustness: Suppose *p* is a semantically approximate product of *p* after being transformed, and *k* is the secret key. If *f* (*p*, *k*) = *f* (*q*, *k*), then *f* is called robust.

Definition 2.3 (information system) The information system is the 4-elements vector R=(*U*, *A*, *V*, *f*), while *U*={ x_1, x_2,…, x_n} is the set of object, A={ a_1, a_2,…, a_n} is the set of the attributes. V= $\bigcup_{a \in A} V_a$ is the range of the A, while *Va* is the range of Feature A.

Definition 2.4 (Entropy and Mutual Information)

The Entropy of the *R* is *L*={*U*,*C*$\cup$*D*,*V*,*F*}， *C* is condition Feature set and *D* is decision Feature set. When $R \subset C$,*Ir* and *Id* drive the devision as: *X*=(x_1, x_2,…, x_n),*Y*=(y_1, y_2,…, y_n), then the entropy of the *R* is as the follow:

$$H(R) = -\sum_{i=1}^{n} p(X_i)\lg X_i\ , \quad p(X_i) = card(X_i)/card(U)\ 。$$

where $p(Y_j \mid X_i) = card(Y_j \cap X_i)/card(X_i)$ 。

Mutual information define: $W(R,D) = H(D) - H(D \mid R)$. (1)

3. Feature extraction algorithm

Feature extraction algorithm *f* makes the important role in the birthmark. The classics VSM(Vector Space Model) is the field of the definition of the Feature extraction algorithm f. VSM simplify the natural language text as the vector space and convey the semantic similarity to space similarity. But the original VSM contains the numerous elements and is too large to be processed. So the reduction of the VSM is necessary to prevent from the large computing cost and low efficiency of the clustering algorithm. From this view, many evaluation methods of the feature importance have been created recently. We proposed a new algorithm based on threshold technology from the viewpoint of information theory as following.

3.1. Threshold Feature Extraction Algorithm

The field of feature a is *a*=(a_1, a_2,…, a_n)，the aim of this algorithm is that random t values in the field can be valid to evaluate the importance of feature *a*. Now we give the formula of the algorithm:

$$J(a,R,D) = -(W(R \cup \{a\},D) - W(R,D))/\varphi(a) \qquad (2)$$

Where $\varphi(a) = d + b_1 x + b_2 x^2 + ... + b_{t-1} x^{t-1}$, *d* is the random integer. Put *ai* into $\varphi(x)$ to get every $\varphi(a_i)$， then the factors (*b1*, *b2*,…, *bt-1*) can be determined using Lagrange's interpolation algorithm.

In the statistics of feature vector, many methods focus on taking into account the correlation and independence of words, but not considering the feasibility of the text. In the establishment of a feature vector space of the text, due to the dilemma of efficiency and compatibility, it is often the case to require repeated tests to finally determine the feature vector dimension involved in the feature assessment. In addition, the content of digital texts duplicated and spread on the Internet may change; therefore, the feature-extract method as a birthmark characterization should have considerable robustness.

3.2. Birthmark constructing algorithm

For the sake of brevity, we have established Birthmark based on the above Threshold Feature Extraction Algorithm. We recall the definition of the Tate pairing [9]. Let r be a positive divisor of the order of $J_c(F_q)$ with $\gcd(r,q)$, and *k* be the smallest integer such that $r \mid (q^k - 1)$; such *k* is called the embedding degree. Let $J_c[r]$ be the divisor classes of order dividing *r*. The Tate *pairing* is a map:

$$e : E[r] \times E(F_{q^k}) / rE(F_{q^k}) \to F_{q^k}^* / (F_{q^k}^*)^r$$

Then the Brithmark algorithm is introduced as following:

```
Algorithm of Brithmark
INPUT： $p$; $a=(a_1,a_2,\ldots, a_n)$;
OUTPUT： $B$=Birthmark of $p$;  $\varphi(X_i)$;
1.Select $a_1, a_2, \ldots, a_t \in a$,
2.Compute $X_i \leftarrow f(a_i)$ while $a_i \in (a_1, a_2, ..., a_n)$;
3.Select $d \in F_q$, $Construct \quad function\varphi \quad as(4)$;
4.$Compute \quad \varphi(X_i) \quad while \quad X_i \in (X_1, X_2, ..., X_t)$;
5.$m = H_2(p) \quad by \quad BelievedThirdParty$;
6.$G_{ID} = H_1(ID) \quad P_{pub} = dQ$;
7.$g_{ID} = e(Q_{ID}, P_{pub})$;
8.$r = H_4(f(a_1), B =< rP, m \oplus H_3(g_{ID}^r) >$;
9$return \quad B$;
```

Since the Attacker modifies some places in the *t* vector of the threshold by lucky coincidence, the output of this algorithm will be distorted to judge the text similarity. We give a heuristic method to use the above algorithm.

3.3 Heuristic method of the birthmark extract algorithm

Firstly, the percent *w* of the similarity should be made. If similarity of text p and q is larger than *w* should be considered the same.

```
INPUT： $p$; $t$; $n$; $a=(a_1,a_2,\ldots, a_n)$;$w$.
OUTPUT： similarity(yes or no) ;
1.$i = 0$
2.$For$    Select $a_{1+i}, a_{2+i}, \ldots, a_{t+i} \in a$,
3.Compute $b \leftarrow Birthmak(a_{1+i}, a_{2+i}, \ldots, a_{t+i})$
4.$Count \quad similarity \leftarrow b$
3.If $similarity > w$
4.    $return \quad yes$
5.$Endfor$
6.$return \quad no$
```

4. Performance analysis

4.1. IND-CCA2 security

Because the Hash function *H* is one-way non-collision, and then the program is IND-CCA2 security.

4.2. Robustness

Firstly, *Attacker* modifies digital works *p* to *q*, attempting to remove or destroy the birthmark *B*, and then sold it to other people. If the robustness of the birthmark is strong, which can resist the transformation by *Attacker*, then the *Provider* is still able to use the secret key *k* select *B* from *q*, which certifies that *q* is the copy of *p*, and then use the copyright verification method to prove his copyright to *q*; If *B* has been completely removed or seriously damaged, then *Provider* cannot select *B*, and as a result, the copyright verification will fail.

For this algorithm, because the (*t*, *n*) threshold program is adopted, the robustness of the birthmark *B* depends the ratio of *t* and *n*. Suppose the value of *n-t* is large enough, then even if the *Attacker* destroyed many features, as long as there are still number *t* features left to the *Provider*, the birthmark is still valid.

4.3. Efficiency

Digital text works always have a numerous field of one feature, such as noun or verb. The most precise method to know the similarity of the p and q of course covers the full field of the feature. But it is impossible to count the numerous elements in the field. So the threshold technology can bring the improvement of the feature-extracting method.

Since Lagrange’s interpolation algorithm is polynomial time, the efficiency of this program depends largely on the feature-selecting algorithm and bilinear pairing of calculation. The feature-selecting algorithm can take various forms, in which the most efficient method is the document frequency algorithm, but the rare words and mutual information are not taken into consideration. Information gain marks features by employing the information theory approaches. Although it is more scientific, its efficiency is the lowest.

5. The copyright program

5.1. Function of copyright program

The functional requirements of the copyright program:

Provider selected the birthmark *B* from his digital work *p* using the secret key, and registered *B* in arbitral institutions and had it added the timestamp *T* by the arbitral institutions. Subsequently, *Provider* sold the *p* to *Consumer*. Three piracy situations happened:

I. *Consumer* sold *p* (which copy named *q*) without any modification to other people without the authorization of the *Provider*. When *Provider* found that he selected the birthmark *B* from *q* using the secret key *k*, thus proving that *q* is a copy of *p*. At the same time, the timestamp *T* registered in arbitral institutions and the secret key *k* certified that *Provider’s* copyright towards *q*.

II. *Consumer* first transformed *p* to generate *q*, attempting to remove or destroy birthmark *B*, and then sold it to other people. If the robustness of birthmark *B* was strong, which could resist the destroying by *Consumer*, then *Provider* was still able to use the secret key *k* to select *B* from *q*, proving *q* was a copy of *p*, and then used the method in situation Ⅰ to prove that the *q*’s copyright belonged to *Provider*; If *B* had been completely removed or seriously damaged, then *Provider* could not select *B*, and the copyright verification failed.

III. *Provider* found that someone was using the similar document *q* in third-party, and verified that it was a copy of *p* with the similar case to Ⅱ. In the case of being unable to certify whether the third-party’s document came from *Consumer* or not, the third party will assume the full legal liability. And if the third party wanted to prove that they did not violate the copyrights of others, they should have to provide the certificate of copyright issued by *Provider*. At this time the source of the text works can be traced down in terms of the certificate.

5.2. Selecting System Parameters

For the establishment of the copyright program of (*t*, *n*) threshold technology, first use the TF-IDF method to establish the feature vector space $C = (C_1, C_2, \ldots, C_n)$. To track down the source of piracy, under normal circumstances we will select a unique feature vector C_k, which can only correspond to one legal Consumer. Therefore, we can divide the vector space $C = (C_1, C_2, \ldots, C_n)$ into two parts: *Consumer Space* = $(C_1, C_2, \ldots, C_m)$, and *Threshold Space* = $(C_{m+1}, C_{m+2}, \ldots, C_n)$, among which $C_k \in$ *Consumer Space*, which is used to mark the *Consumer*, while the *Threshold Space* is used for the establishment of threshold system.

Meanwhile, in order to improve the practicality of the system, taking into consideration of improving the system based on identity authorization function, we adopting the encryption program using bilinear matching technology as a birthmark. For the bilinear matching technology, the efficiency of Tate pairings is higher than Weil pairings [9,12], and later we can adopt a variety of technology such as Eta, Ate, etc. to accelerate pairings. therefore, this program uses the Tate to achieve it.

The Algorithm of Copyright Release

```
1.Select $\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_t \in \mathrm{C}$, while $\mathrm{C}_1 \in$ Consumer Space, $\mathrm{C}_2, \ldots, \mathrm{C}_t \in$ Threshold Space;
2. Compute $\mathrm{X}_i \leftarrow f(\mathrm{C}_i)$ while $\mathrm{C}_i$�$(\mathrm{C}_1, \mathrm{C}_2, ..., \mathrm{C}_n)$;
3. Select $\mathrm{d} \in \mathrm{F}_q$, $Construct \quad function\, \varphi \quad as(4)$;
4.$Compute \quad \varphi(X_i) \quad while \quad X_i \in (X_1, X_2, ..., X_t)$;
5.$m = H_2(p) \quad by \quad BelievedThirdParty$;
6.$G_{ID} = H_1(ID) \quad \mathrm{P_{pub}} = \mathrm{dQ}$;
7.$g_{ID} = e(Q_{ID}, P_{pub})$;
8.$r = H_4(f(C_1), B =< rP, m \oplus H_3(g_{ID}^r) >$;
9$return \quad B$;
```

Procedure R(*p*,*C*,*ID*)

INPUT：*p*; *C*=(*C1, C2,…, Cn*); *Consumer ID*;

OUTPUT：*B*=Birthmark of *p*; $\varphi(X_i)$;

Provider transmits the *dID* = *dGID* to Consumer through the security message channel. Later, after Consumer receives B, he will get the copyright information through $H_3(g_{ID}^r) \oplus H_3(e(d_{ID}, rP)) = m$

Copyright Verification Algorithm

Procedure V(*B*,*C*,*ID*)

INPUT：*B*; *C*=(*C1, C2,…, Cn*); *Consumer ID*;

OUTPUT：*B* is valid or not;

```
1.Select $\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_t \in \mathrm{C}$ and compute d by them;
2. Compute m from B and d;
3 If m is valid return TRUE, else return FALSE.
```

Algorithm of Copyright Retrospect

Provider found that one user is using a text file *q*, which is very similar to *p*. *Provider* can trace down its copyright, and can demand the copyrighted file from that user. If the user can not provide it, it will be proved that he violates the copyright; if he can provide the copyright data, then it can be done to track down the original user through the copyright data.

Algorithm of Copyright Retrospect

Procedure R(*B*,*C*,*ID*)

INPUT：*B*;*C1*;

OUTPUT：Whether *B* is judged by *C*1;

```
1.$If \quad H_4(\mathrm{C}_1)\mathrm{P} = U \quad While\, B(U,V)$
2       return TRUE;
3 else
4       return FALSE;
```

6. Conclusion

We proposed a threshold (*t*,*n*) birthmark selecting technology towards digital-text works in this paper, which can freely adjust the vector space dimension birthmark. This technique will take up and resist the data loss problem when text works are transmitted or maliciously modified on the Internet. Based on this technology, a copyright program of the digital-text works was put forward to and endued with the function of tracing down source of release and IND-CCA2 security. In the achieving process, there is still great place for improving the efficiency by using bilinear pairings. Therefore, the further task is focusing on improving the efficiency of this program through a variety of techniques.

References

[1]. Tamada H, Nakamura M, Monden A, et al. Design and Evaluation of Birthmarks for Detecting Theft of Java Programs. Proc. of IASTED'04. Spain: ACTA Press, 2004: 569-575.

[2]. Tamada H, Nakamura M, Monden A, et al. Java Birthmarks—Detecting the Software Theft. IEICE Transactions on Information and Systems, 2005, E88-D(9): 2148-2158.

[3]. Vassaux B, Bas P,and Chassery J M. A new CDMA technique for digital image watermarking, enchancing capacity of insertion and robustness[A] .Proc.of IEEE Int. conf. on Image processing. Thessalonica, Greece. 2001, 3:983-986.

[4]. Frank Hartung, Mrtin Kutter. Multimedia Watermarking Techniques .Proceeding of the IEEE, VOL.87(7), 1079-1107,July. 1999.

[5]. Salton G.Term weighting approaches in automatic text retrieval. Information processing & management.1988 / 24 / 05 P 513-523.

[6]. Yi-Ming Yang, Jan O Pederson. A Comparative Study on Feature Selection in Text Categorization. Proc. of 14 th International Conference on Machine Learning (ICML - 97), 1997, 412 - 420.

[7]. Marcus Hutter**,** Marco Zaffalon. Distribution of mutual information for robust feature selection. Tech. rept. IDSIA-11-02. IDSIA, Manno (Lugano), CH. Submitted.

[8]. Hutter, M.. Distribution of Mutual Information. Proceedings of the 14th International Conference on Neural Information Processing Systems (NIPS-2001). In press.

[9]. Dan Boneh, Mathhew Franklin. Identity-Based Encryption From Weil Pairing. http://eprint.iacr.org/.

[10]. Pedersen T. Non-interactive and information theoretic secure verifiable secret sharing. Advances in Cryptology-Crypto'91. 1991.129-140.

[11]. Chor B, Goldwasser S, Micali S, Awerbuch B. Verifiable secret sharing and achieving simultaneity in the presence of faults. Proceedings of 26th IEEE symposium on foundations of computer science 1985.251-160.

[12]. Neal Koblitz, Alfred Menezes, Pairing-based cryptography at high security levels. http://eprint.iacr.org/.

[13]. P.S.L.M. Barreto, S. Galbraith, M. Scott. Effcient Pairing Computation on Supersingular Abelian Varieties. Design, Codes and Cryptography, Vol. 42,No.3, pp.239-271, 2007.