Kỷ yếu Hội nghị Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00049

MỘT LƯỢC ĐỒ THỦY VÂN THUẬN NGHỊCH KHÓA CÔNG KHAI CHO CƠ SỞ DỮ LIỆU DỰA TRÊN KỸ THUẬT MỞ RỘNG CÁC THUỘC TÍNH KIỂU THỰC

Lê Quang Hòa[1], Đỗ Văn Tuấn[2], Nguyễn Huy Đức[3], Phạm Văn Ất[4]

[1] Viện Toán ứng dụng và Tin học, Đại học Bách khoa Hà Nội
[2] Khoa Công nghệ thông tin, Đại học Công nghiệp Hà Nội
[3]Cao Đẳng Sư phạm Trung ương
[4]Đại học Giao thông Vận tải

hoa.lequang1@hust.edu.vn, dvtuanhn@gmail.com, ducnghuy@gmail.com, phamvanat83@vnn.vn

TÓM TẮT— *Thủy vân số được xem là một công cụ hữu hiệu để bảo vệ cơ sở dữ liệu trên các môi trường trao đổi không an toàn. Các lược đồ thủy vân truyền thống chỉ cho phép trích thủy vân mà không có khả năng khôi phục cơ sở dữ liệu. Do đó, người nhận không có được dữ liệu gốc mà chỉ là một bản gần đúng. Để khắc phục nhược điểm trên, gần đây đã xuất hiện các lược đồ có khả năng khôi phục được bản gốc, gọi là các sơ đồ thủy vân thuận nghịch. Phép biến đổi mở rộng hiệu được xem là một kỹ thuật hữu hiệu để xây dựng các lược đồ thủy vân thuận nghịch, tuy nhiên nó có nhược điểm làm cho dữ liệu bị biến đổi khá nhiều. Gần đây Fafoura và cộng sự đề xuất sử dụng phương pháp mở rộng hiệu trên phần phân của các thuộc tính kiểu số thực, do đó giảm sự thay đổi và nâng cao chất lượng thủy vân. Tuy nhiên bằng các phân tích lý thuyết cũng như thử nghiệm, chúng tôi chỉ ra trong một số trường hợp lược đồ này không thể khôi phục thủy vân cũng như dữ liệu gốc một cách chính xác. Trong bài báo này chúng tôi đề xuất một giải pháp khắc phục nhược điểm nói trên bằng cách sử dụng bản đồ định vị để phân biệt trường hợp có thể nhúng theo phương pháp mở rộng phần phân và trường hợp không cho phép nhúng. Sau đó dựa trên phương pháp nhúng tin mới này, chúng tôi đề xuất một lược đồ thủy vân thuận nghịch dễ vỡ khóa công khai trên các cơ sở dữ liệu quan hệ. Lược đồ này có ưu điểm ở chỗ người nhận dễ dàng kiểm tra tính toàn vẹn của cơ sở dữ liệu thủy vân nhận được và khôi phục đúng cơ sở dữ liệu gốc. Như vậy người nhận không phải sử dụng cơ sở dữ liệu gần đúng (thủy vân) mà dùng cơ sở dữ liệu chính xác.*

Từ khóa— *Cơ sở dữ liệu quan hệ, thủy vân thuận nghịch, mở rộng hiệu, dễ vỡ, khóa công khai.*

I. GIỚI THIỆU

Thủy vân là kỹ thuật nhúng thêm thông tin vào dữ liệu đa phương tiện như hình ảnh, âm thanh và cơ sở dữ liệu, trước khi những dữ liệu này được trao đổi trên môi trường không an toàn. Các thuật toán nhúng thông tin (dấu thủy vân) thường biến đổi dữ liệu gốc theo một qui tắc nào đó để nhận được dữ liệu chứa tin (dữ liệu thủy vân), nên giữa chúng luôn có một sự sai lệch nhất định. Tuy nhiên, dấu thủy vân là cơ sở để phát hiện những sự thay đổi trái phép trong quá trình trao đổi, hoặc dùng để chứng minh quyền sở hữu khi có sự tranh chấp.

Theo [1,8-10,20], các lược đồ thủy vân được chia thành thủy vân dễ vỡ và thủy vân bền vững. Thủy vân dễ vỡ [2-12] yêu cầu dấu thủy vân phải nhạy cảm (dễ vỡ) trước những sự tấn công trái phép trên dữ liệu thủy vân, dù chỉ vài bít. Do đó các lược đồ này thường được dùng trong việc phòng chống giả mạo, hay xác thực tính toàn vẹn của dữ liệu thủy vân. Trái với thủy vân dễ vỡ, các lược đồ thủy vân bền vững [16,21] lại mong muốn dấu thủy vân ít bị biến đổi (bền vững) trước các phép tấn công. Vì vậy thủy vân bền vững được dùng trong bài toán bảo vệ bản quyền.

Mặt khác theo [1,20], các lược đồ thủy vân có khả năng khôi phục lại dữ liệu gốc từ dữ liệu thủy vân thì được gọi là thủy vân thuận nghịch (reversible watermarking). Thủy vân thuận nghịch thường sử dụng một số phép biến đổi như: mở rộng hiệu [2,7-8,10-11, 14-15,19]; dịch chuyển histogram [6]; wavelet nguyên [12], nén bảo toàn [18,22] và sử dụng đặc trưng nén JPEG [3-4,9]. Ngoài ra, việc kết hợp kỹ thuật dự báo với mở rộng hiệu hoặc dịch chuyển histogram cũng nhận được nhiều sự quan tâm của các nhà nghiên cứu. Theo các tài liệu [1,20], trong nhiều ứng dụng của y tế, quân sự và nghệ thuật, việc khôi phục lại ảnh gốc từ ảnh thủy vân là một trong những yêu cầu bắt buộc. Bởi, chỉ cần một sự thay đổi nhỏ trên ảnh sử dụng so với ảnh gốc cũng có thể ảnh hưởng đến kết quả chuẩn đoán của bác sỹ, hoặc gây ra các hệ lụy nghiêm trọng trong quân sự.

Để bảo vệ cơ sở dữ liệu quan hệ, trong [21] đề xuất nhúng dãy bít thủy vân vào bít thấp của các thuộc tính số, những thuộc tính chấp nhận sự thay đổi nhỏ nhưng không ảnh hưởng đến ý nghĩa của dữ liệu. Theo [21], bộ dữ liệu cũng như thuộc tính dùng để nhúng dấu thủy vân được chọn ngẫu nhiên và bí mật thông qua mã hàm băm ứng với giá trị của trường khóa chính. Đây được xem là lược đồ thủy vân đầu tiên trên cơ sở dữ liệu quan hệ. Để cải thiện tính bền vững, trong [16-17] đề xuất nhúng nhiều bít thủy vân vào mỗi thuộc tính được chọn thay cho nhúng một bít như [21].

Dựa trên ý tưởng của [16,17,21], trong [13] đề xuất lược đồ thủy vân dễ vỡ bằng cách sử dụng khóa bí mật để phân hoạch các bộ dữ liệu (các bản ghi) của quan hệ $R = (\mathrm{P}, A_0, \ldots A_{n-1})$ thành các tập con, trong đó P là khóa chính của quan hệ R và A_i là các thuộc tính số,và nhúng dấu thủy vân vào tất cả các bộ dữ liệu trong các tập con theo phương pháp chèn bít thấp. Kết quả thực nghiệm trong [13] cho thấy, lược đồ này có khả năng phát hiện được nhiều dạng tấn công khác nhau trên cơ sở dữ liệu thủy vân. Tuy nhiên, việc thay đổi bít thấp của các trường có kiểu số nguyên trên tất cả các bản ghi như [13] sẽ tạo ra một sự sai lệch không nhỏ giữa dữ liệu thủy vân so với dữ liệu gốc. Mặt khác, các

lược đồ [13, 16,17,2] không có khả năng khôi phục lại cơ sở dữ liệu gốc, điều này đồng nghĩa với việc người dùng phải sử dụng dữ liệu sai lệch so với dữ liệu gốc.

Để giảm sự sai lệch giữa dữ liệu thủy vân so với dữ liệu gốc, trong [10] (lược đồ Farfoura) đề xuất nhúng bít thủy vân vào phần phân của các trường có kiểu số thực theo phương pháp mở rộng hiệu. Do vậy về lý thuyết, lược đồ [10] có khả năng khôi phục lại dữ liệu gốc từ dữ liệu thủy vân. Tuy nhiên trong thực tế, lược đồ [10] có sự nhầm lẫn dẫn đến trường hợp không thể khôi phục chính xác các bít thủy vân cũng như giá trị các thuộc tính dữ liệu gốc. Điều này sẽ được phân tích trong Mục II.B. Dựa trên ý tưởng [10], bài báo này đề xuất lược đồ thủy vân thuận nghịch dễ vỡ trên cơ sở dữ liệu quan hệ bằng cách sử dụng một bản đồ định vị để phân biệt khi nào có thể nhúng tin thuận nghịch và khi nào không thể làm điều đó. Các phân tích và kết quả thực nghiệm cho thấy, lược đồ đề xuất có khả năng phát hiện được nhiều dạng tấn công trái phép trên cơ sở dữ liệu thủy vân.

Nội dung tiếp của bài báo được tổ chức như sau: Mục II giới thiệu một số công trình liên quan. Mục III trình bày phương pháp nhúng tin đề xuất. Mục IV xây dựng một lược đồ thủy vân thuận nghịch khóa công khai. Kết quả thực nghiệm trình bày trong Mục V. Cuối cùng là một số kết luận trong Mục VI.

II. NHỮNG CÔNG TRÌNH LIÊN QUAN

Phần này trình bày một số lược đồ thủy vân thuận nghịch liên quan đến phương pháp đề xuất.

A. Phương pháp mở rộng hiệu

Biến đổi mở rộng hiệu được đề xuất bởi Tian [14] dùng để xây dựng lược đồ thủy vân trên ảnh số. Theo [14], việc nhúng bít b vào cặp điểm ảnh (x,y) có giá trị thuộc [0,255] thực hiện theo công thức:

$$h = x - y, \qquad l = \left\lfloor \frac{x+y}{2} \right\rfloor \tag{2.1}$$

$$h' = 2h + b, \qquad x' = l + \left\lfloor \frac{h'+1}{2} \right\rfloor, \qquad y' = l - \left\lfloor \frac{h'}{2} \right\rfloor \tag{2.2}$$

Khi đó, việc trích bít b và khôi phục cặp điểm ảnh (x,y) từ cặp điểm ảnh thủy vân (x', y') thực hiện theo các công thức:

$$h' = x' - y', \qquad l = \left\lfloor \frac{x'+y'}{2} \right\rfloor, h = \left\lfloor \frac{h'}{2} \right\rfloor$$

$$b = h' \bmod 2, \quad x = l + \left\lfloor \frac{h+1}{2} \right\rfloor, \qquad y = l - \left\lfloor \frac{h}{2} \right\rfloor$$

Nếu sau khi nhúng tin, các giá trị $x' và y'$ vẫn nằm trong miền giá trị của điểm ảnh (miền này bằng [0,255] đối với ảnh đa cấp xám), thì có thể nhúng thuận nghịch một bít trên cặp điểm ảnh (x, y). Để phân biệt cặp nào có thể nhúng, cặp nào không thể nhúng, Tian đã sử dụng một bản đồ định vị, thực chất là một dãy nhị phân có giá trị 0 hoặc 1 ứng với cặp có thể và không thể nhúng tin. Độ dài bản đồ bằng một nửa số điểm ảnh.Vì bản đồ định vị cần dùng trong việc khôi phục thủy vân và ảnh gốc, nên nó được nén và nhúng vào ảnh gốc cùng với dãy bit thủy vân. Điều này làm giảm đáng kể khả năng nhúng của phương pháp mở rộng hiệu.

B. Thủy vân thuận nghịch trên cơ sở dữ liệu

Dựa trên ý tưởng của Tian [14], lược đồ Gupta [11] đề xuất nhúng một bít thủy vân trên hai thuộc tính kiểu số nguyên của cơ sở dữ liệu quan hệ ứng với một bản ghi nào đó thông qua phép biến đổi mở rộng hiệu. Hạn chế chính của lược đồ Gupta là dữ liệu sau khi nhúng thủy vân có sự biến đổi khá lớn so với dữ liệu gốc.

Nhằm nâng cao chất lượng dữ liệu thủy vân, các tác giả trong [10] (lược đồ Farfoura) đề xuất nhúng một bít thủy vân vào phần phân của một thuộc tính số thực theo biến đổi mở rộng hiệu. Theo [10], việc nhúng bít b vào số thực X để nhận được X' thực hiện theo các bước:

- Tách X thành phần nguyên n và phần phân p: [n,p]=Tach(X)
- Nhúng bít b vào phần phân p theo công thức: p'=2*p+b
- Ghép phần nguyên n với phần phân mới p' để tạo X': X'=Ghep(n,p')

Ví dụ nếu b=1 và X=3.18 thì n=3, p=18, p'=2*p+b=37 và X'=Ghep(3,37)=3.37.

Để trích bít b và khôi phục X từ X' ta thực hiện theo trình tư ngược lại:

- Tách n và p' từ X': [n,p']=Tach(X')
- Khôi phục b và p từ p': b=p' mod 2, p=p' div 2
- Ghép n và p để được X: X=Ghep(n,p)

Ví dụ với X'=3.37 thì n=3, p'=37, nên b= 37 mod 2=1, p=p' div 2=18 và X=Ghep(n, p)=3.18

Do chỉ biến đổi trên phần phân của những thuộc tính kiểu số thực, nên lược đồ Farfoura có ưu điểm ở chỗ sự thay đổi trong cơ sở dữ liệu gốc không nhiều. Tuy nhiên do quy tắc không lưu trữ các chữ số 0 tận cùng bên phải của

phần phân, nên nhiều trường hợp lược đồ Farfoura không thể trích đúng bít thủy vân và khôi phục được dữ liệu gốc, đây có thể được xem là lỗi khá nghiêm trọng trong thủy vân thuận nghịch.

Để dễ hình dung sự nhầm lẫn trong [10], ta xét ví dụ X =3.145 và bít thủy vân b =0, khi đó thuật toán nhúng tin của lược đồ Farfoura sẽ cho kết quả $X' = 3.290$, nhưng các hệ quản trị CSDL chỉ lưu trữ giá trị 3.29 thay vì 3.290. Khi đó, thuật toán trích tin và khôi phục dữ liệu của lược đồ Farfoura ứng với $X' = 3.29$ sẽ là b=1 và $X' = 3.14$. Như vậy, trong trường hợp này lược đồ Farfoura là không chính xác. Điều này có thể khắc phục được bằng cách sử dụng một bản đồ định vị để phân biệt trường hợp có thể và không thể nhúng tin. Giải pháp này được trình bày chi tiết trong phương pháp đề xuất.

III. PHƯƠNG PHÁP ĐỀ XUẤT

A. Ý tưởng của phương pháp

Trước tiên ta khảo sát sự biến đổi chữ số đơn vị của p (gọi cho gọn là đuôi và ký hiệu D(p)) qua phép biến đổi p'=2*p+b. Sự biến đổi này dễ dàng được biểu diễn qua bảng sau:

Bảng 1. Khảo sát sự biến đổi của D(p) qua phép biến đổi p'=2*p+b

TT	D(p)	D(p')	
		b=0	b=1
1	0(số nguyên)	0	1
2	1	2	3
3	2	4	5
4	3	6	7
5	4	8	9
6	5	0	1
7	6	2	3
8	7	4	5
9	8	6	7
10	9	8	9

Theo bảng 1 trường hợp b=0 và D(p)=5 thì D(p')=0. Như đã phân tích ở trên trường hợp này sẽ dẫn đến lỗi trích bít b và khôi phục phần phân p. Như vậy khi gặp trường hợp này thì cần bỏ qua và không nhúng bít thủy vân.

Để biết khi nào có thể và khi nào không thể nhúng tin, ta phân hoạch tập đuôi p {0, 1,.., 8, 9} thành hai tập rời nhau N và K. Khi D(p) thuộc N thì ta có thể nhúng thuận nghịch một bít b vào giá trị p để nhận p', còn khi D(p) thuộc K thì p không được dùng để nhúng tin. Ta gọi (N, K) là bản đồ định vị để phân biệt các trường hợp nhúng và không nhúng. Như sẽ chỉ ra dưới đây, khi biết p' có thể suy ra D(p) thuộc N hay K, do đó không cần nén và nhúng thông tin về bản đồ vào cơ sở dữ liệu gốc như trong phương pháp của Tian nên khả năng nhúng tin của phương pháp đề xuất khá cao.

Bây giờ sẽ xác định bản đồ (N, K). Như phân tích ở trên, 5 thuộc K và khi D(p)=5 thì ta không nhúng tin và đặt p'=p=5. Theo như bảng 1, khi D(p)=2 hoặc D(p)=7 và b=1 thì D(p')=5. Như vậy trong trường hợp này ta cũng không thể nhúng tin. Nói cách khác, 2 và 7 cũng phải thuộc K. Cũng theo bảng 1, các trường hợp còn lại đều có thể sử dụng để nhúng tin, vậy có thể xác định bản đồ (N, K) như sau:

K={2, 5, 7} và N={0, 1, 3, 4, 6, 8, 9}

Sau khi đã xác định được bản đồ (N, K) có thể dễ dàng xây dựng phương pháp nhúng 1 bít thủy vân trên 1 giá trị thuộc tính kiểu số thực như trình bày dưới đây.

B. Phương pháp nhúng trên một giá trị của thuộc tính kiểu số thực

1. Thuật toán nhúng 1 bít thủy vân

Đầu vào: X=t.A- giá trị bộ thứ t thuộc tính kiểu thực A, b- bít nhúng giá trị 0 hoặc 1

Đầu ra: X'=t.A' giá trị mới bộ thứ t thuộc tính kiểu thực A

Bước 1: Từ giá trị thực X tách thành 3 phần: phần nguyên n, số chữ số không sau phần nguyên s và phần phân p

[n, s, p]=Tach(X)

Ví dụ nếu X=24.000567 thì n=24, s=3, p=567.

Bước 2: Dựa vào D(p) và bản đồ (N, K) để phân biệt trường hợp nhúng và không nhúng, cụ thể như sau:

Trường hợp 1: nếu $D(p) \in K$ không nhúng chuyển đến bước 3

Trường hợp 2: nếu $(p) \in N$, nhúng, chuyển đến bước 4

Bước 3: Xử lý trường hợp không nhúng tin

Trường hợp 1: nếu D(p)=5 thì giữ nguyên p: p'=p

Trường hợp 2: nếu $D(p) \in \{2,7\}$ thì biến đổi p theo công thức $p' = 2 * p$

Chuyển xuống bước 5

Bước 4: Nhúng bít b vào p theo công thức

$p' = 2 * p + b$

Bước 5: Xác định X' bằng cách ghép n, s và p': X'=Ghep(n, s, p')

Nhận xét 1 (tính chất của thuật toán nhúng tin):

Qua các bước trên dễ dàng thấy rằng thuật toán nhúng tin có tính chất sau:

- Nếu D(p) =5 thì D(p')=5
- Nếu $D(p) \in \{2,7\}$ thì D(p')=4
- Nếu D(p) ∈ N thì D(p') ∈ {0,1, 2, 3, 6, 7, 8, 9}

Như vậy từ p' thì có thể suy ra D(p) thuộc K hoặc N và trong trường D(p) thuộc K có thể suy ra D(p) =5 hoặc D(p) thuộc {2, 7}

Từ nhận xét trên, chúng ta xây dựng thuật toán trích bít thủy vân b và khôi phục giá trị X từ X' như sau:

2. Thuật toán khôi phục:

Đầu vào: X'= t.A'- giá trị bộ thứ t thuộc tính kiểu thực đã nhúng hoặc không nhúng bit b

Đầu ra: b- bít nhúng nếu có, X=t.A giá trị thuộc tính kiểu thực gốc

Bước 1: Từ X' tách phần nguyên n, số chữ số không trước phần phân s và phần phân p':

[n, s, p']=Tach(X')

Bước 2: Dựa vào D(p') để phân biệt trường hợp có nhúng tin như sau:

Trường hợp 1: nếu $D(p') \in \{4,5\}$ không trích tin, chuyển đến bước 3

Trường hợp 2: nếu $D(p') \in \{0,1,2,3,6,7,8,9\}$ trích tin, chuyển đến bước 4

Bước 3: Khôi phục p từ p'

Trường hợp 1: nếu D(p')=5 thì p=p'

Trường hợp 2: nếu D(p)=4 thì p=p' div 2

Chuyển xuống bước 5

Bước 4: trích bít b và khôi phục p từ p'

b=p' mod 2

p= p' div 2

Bước 5: Khôi phục giá trị X=t.A bằng cách ghép các thành phần n, s và p:

X=Ghep(n,s,p)

Nhận xét 2 (tính đúng đắn của thuật toán khôi phục): Tính đúng đắn của thuật toán khôi phục có thể dễ dàng chứng minh bằng cách sử dụng tính chất của thuật toán nhúng tin phát biểu trong nhận xét 1.

Một số ví dụ minh họa thuật toán nhúng tin và thuật toán khôi phục

Ví dụ 1: X=21.365 thì n=21, s=0, p=365, không nhúng tin, p'=p=365 và X'=X=21.365.

Ví dụ 2: X=30.017 thì n=30, s=1, p=17, không nhúng tin, p'=2*p=34 và X'=30.034.

Ví dụ 3: X=25.28 và b=1 thì n=25, s=0, p=28, nhúng tin, p'=2*p+b=57 và X'=25.57.

Ví dụ 4: X=18 và b=1 thì n=18, s=0, p=0, nhúng tin, p'=2*p+b=1 và X'=18.1.

C. Phương pháp nhúng tin trên cơ sở dữ liệu quan hệ

1. Thuật toán nhúng

Đầu vào cơ sở dữ liệu quan hệ R= R_ω(P, A_1 , A_2,... A_α, B_1 , B_2,... B_β) trong đó P là thuộc tính khóa, A_1 , A_2,... A_α các thuộc tính giá trị thực, B_1 , B_2,... B_β các thuộc tính kiểu khác, ω số bộ (bản ghi) trong bảng dữ liệu, W=(w_1 , w_2,... w_γ) là dãy γ bít thủy vân

Đầu ra là cơ sở dữ liệu R' chứa W

Bước 1: Xác định khả năng nhúng (số bít tối đa có thể nhúng) trên cơ sở dữ liệu R

Ký hiệu t.A_i là giá trị thuộc tính A_i của bộ thứ t, p_i là phần phân của t.A_i. Nếu đuôi D(p_i) thuộc tập N, thì theo mục A, ta có thể nhúng 1 bít thủy vân trên giá trị t.A_i. Khi đó ta gọi t.A_i là khả nhúng.

Gọi σ_i là số giá trị khả nhúng trong số các giá trị t.A_i với t=1,..., ω, khi đó: $\sigma = \sum_{i=1}^{\alpha} \sigma_i$ là khả năng nhúng trên cơ sở dữ liệu R

Bước 2: Điều chỉnh dãy bít thủy vân W theo khả năng nhúng σ

Nếu số bít thủy vân γ nhỏ hơn hoặc bằng σ thì giữ nguyên W, trái lại ta ngắt bỏ $\gamma - \sigma$ bít cuối của W. Nói cách khác, W chỉ còn σ bít đầu tiên: W=(w_1 , w_2,... w_σ). Như vậy sau khi điều chỉnh, dãy W có độ dài $\theta = \min(\sigma, \gamma)$: W=($w_1$, w_2,... w_θ)

Bước 3: Nhúng W vào cơ sở dữ liệu R

Chọn θ giá trị khả nhúng bất kỳ của R, gọi các giá trị này là X_1 , X_2,... X_θ. Nhúng mỗi bít w_i vào X_i để được X'_i theo thuật toán B1

Bước 4: Tạo cơ sở dữ liệu thủy vân R'

R' được tạo từ R sau khi thay các giá trị X_i bằng X'_i

2. Thuật toán khôi phục W và R

Đầu vào là cơ sở dữ liệu thủy vân R'. Ngoài ra biết tập các giá trị X'={X'_1 , X'_2,... X'_θ} là các giá trị thủy vân nhận được trong thuật toán nhúng.

Đầu ra là dãy bít thủy vân W=(w_1, w_2,... w_θ) và cơ sở dữ liệu gốc R.

Bước 1: Từ dãy X' trích dãy bít thủy vân W và khôi phục dãy giá trị X={X_1 , X_2,... X_θ} theo thuật toán III.B.2

Bước 2: Khôi phục cơ sở dữ liêu gốc R: R được tạo từ R' sau khi thay các giá trị X'_i bằng X_i

Nhận xét 3 (về khả năng nhúng): ở trên đã chỉ ra số bít có thể nhúng được vào thuộc tính A_i bằng số các giá trị khả nhúng trong số ω giá trị t.A_i với i =1,.., ω. Ngoài ra, t.A_i được gọi là khả nhúng nếu phần phân p_i của nó có đuôi thuộc tập N. Do tập N có 7 chữ số, trong khi tập đầy đủ các đuôi là 10, vì vậy nếu các đuôi xuất hiện đều nhau thì số giá trị khả biến xấp xỉ bằng 70% tổng số giá trị. Như vậy có thể thấy khả năng nhúng tin của phương pháp đề xuất bằng 70% số giá trị của các thuộc tính kiểu số thực trong cơ sở dữ liệu, và có thể biểu diễn theo công thức sau:

$$\mathcal{C} = \frac{70 * \alpha * \omega}{100}$$

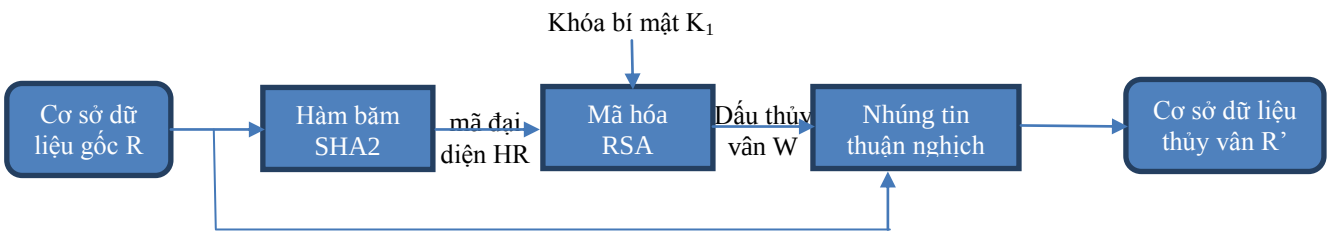
IV. LƯỢC ĐỒ THỦY VÂN THUẬN NGHỊCH DỄ VỠ KHÓA CÔNG KHAI

Dựa vào phương pháp đề xuất ở trên chúng tôi xây dựng một lược đồ thủy vân thuận nghịch dễ vỡ khóa công khai áp dụng cho việc xác thực tính toàn vẹn của cơ sở dữ liệu như sau:

Giả sử cơ sở dữ liệu quan hệ có bảng R=R(P, A_1 , A_2,... A_α, B_0 , B_1,... B_β) trong đó P là thuộc tính khóa A_i là các thuộc tính số thực, B_i là các thuộc tính kiểu khác. Dữ liệu thủy vân được nhúng vào phần phân của giá trị các thuộc tính kiểu số thực A_1 , A_2,... A_α.

A. Thuật toán nhúng thủy vân

Thuật toán tạo và nhúng dấu thủy vân trên cơ sở dữ liệu gốc R để được cơ sở dữ liệu thủy vân R' thực hiện theo sơ đồ sau:



Hình 1. Sơ đồ nhúng thủy vân thuận nghịch dễ vỡ khóa công khai cho cơ sở dữ liệu

Trên Hình 1, thuật toán sử dụng hàm băm SHA2-348 (ký hiệu SHA2) để xác định mã đại diện cho cơ sở dữ liệu R. Khi đó, mã đại diện HR có độ dài 348 bít và được mã hóa bằng hệ mật mã RSA thông qua khóa bí mật K_1 để nhận được dấu thủy vân W. Dấu thủy vân W được nhúng vào cơ sở dữ liệu gốc bằng thuật toán đề xuất để nhận được cơ sở dữ liệu thủy vân R'. Cơ sở dữ liệu thủy vân được dùng để trao đổi trên các kênh truyền công khai.

Chi tiết thuật toán được thể hiện qua các bước sau:

Đầu vào là cơ sở dữ liệu gốc R

Đầu ra là cơ sở dữ liệu thủy vân R’

Bước 1: Xác định mã đại diện HR

HR=SHA2(R)

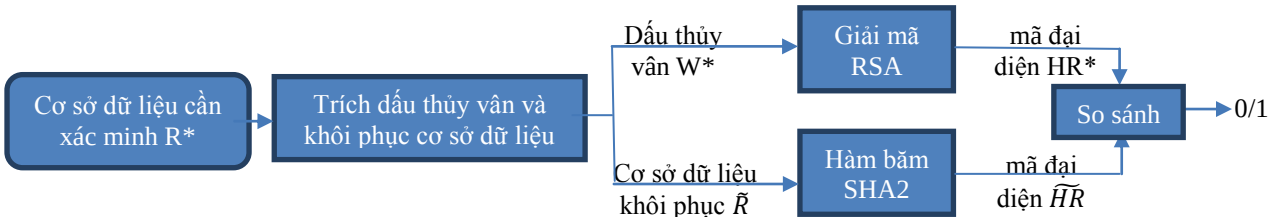
Bước 2: Mã hóa HR bằng mật mã RSA dùng khóa bí mật K_1

W=MaHoaRSA(HR, K_1)

Bước 3: Nhúng W vào cơ sở dữ liệu R theo thuật toán III.C.1 được cơ sở dữ liệu thủy vân R’

B. Thuật toán xác thực tính toàn vẹn và khôi phục cơ sở dữ liệu gốc

Cơ sở dữ liệu thủy vân R’ được chuyển đến người nhận qua môi trường công khai và có thể bị biến đổi thành R*. Do đó, người nhận có được R* và cần xác minh xem đây có đúng là cơ sở dữ liệu thủy vân R’ hay không. Thuật toán xác thực được mô tả theo sơ đồ dưới đây:



Hình 2. Sơ đồ xác thực tính toàn vẹn thủy vân thuận nghịch dễ vỡ khóa công khai cho cơ sở dữ liệu

Trên hình 2 đầu tiên trích dấu thủy vân W* và khôi phục cơ sở dữ liệu $\tilde{R}$ từ R* bằng thuật toán đề xuất. Dấu thủy vân được giải mã RSA với khóa công khai K_2 để nhận được mã đại diện HR*. Mặt khác, từ cơ sở dữ liệu vừa khôi phục sử dụng hàm băm SHA2 để lấy mã đại diện $\widetilde{HR}$. Ta so sánh $\widetilde{HR}$ và HR* nếu chúng giống nhau kết luận R’* là cơ sở dữ liệu thủy vân và R* chính là cơ sở dữ liệu gốc. Ngược lại kết luận cơ sở dữ liệu R’ đã bị tấn công trên đường truyền.

Chi tiết thuật toán được thể hiện qua các bước sau:

Đầu vào cơ sở dữ liệu cần xác thực R*

Đầu ra là kết quả xác thực xem R* có đúng là R’ hay không. Nếu đúng thì khôi phục cơ sở dữ liệu gốc R.

Bước 1: Trích dấu thủy vân W* và khôi phục cơ sở dữ liệu $\tilde{R}$ từ R*

Bước 2: Xác định mã đại diện theo hai cách

HR*=GiaiMaRSA(W*,K_2)

$\widetilde{HR}$=SHA2(R*)

Bước 3: Xác thực tính toàn vẹn

Nếu $\widetilde{HR}$=HR* thì dữ liệu toàn vẹn, có nghĩa là R*=R’ và $\tilde{R}$ =R

Ngược lại dữ liệu không toàn vẹn, cơ sở dữ liệu đã bị tấn công(R*≠R’).

C. Về tính đúng đắn của mô hình đề xuất

Do dấu thủy vân là đại diện của toàn bộ cơ sở dữ liệu và được xác định bởi hàm băm nên chỉ cần một sự thay đổi nhỏ dù chỉ một giá trị thuộc tính của bản ghi nào đó thì mã đại diện cũng bị biến đổi (tính chất của hàm băm). Vì vậy, mô hình thủy vân đề xuất có khả năng phát hiện mọi sự tấn công trái phép trên cơ sở dữ liệu đã thủy vân. Điều này phù hợp với kết quả thực nghiệm.

V. THỬ NGHIỆM

Để thử nghiệm khả năng phát hiện các dạng tấn công lên cơ sở dữ liệu thủy vân R’, chúng tôi sử dụng cơ sở dữ liệu gốc R có một bảng dữ liệu gồm bốn thuộc tính thực và 5000 bản ghi. Sau mỗi lần tấn công, thuật toán tính các giá trị đại diện HR* và $\widetilde{HR}$, chúng đều là các dãy 384 bít. Để so sánh sự khác nhau giữa HR* và $\widetilde{HR}$, chúng tôi sử dụng độ sai khác bình phương trung bình:

$$\rho(HR^*, \widetilde{HR}) = \frac{\sum_{i=1}^{384}(HR_i^* - \widetilde{HR}_i)^2}{384}$$

Nếu $\rho(HR^*, \widetilde{HR}) = 0$ thì có thể kết luận R’ không bị tấn công (R*=R’), trái lại thì kết luận bị tấn công (R*≠R’).

Kết quả thử nghiệm trình bày trong bảng dưới đây cho thấy mọi sự thay đổi dù nhỏ đối với cơ sở dữ liệu thủy vân đều được phát hiện.

Bảng 2. Thử nghiệm sự khác nhau giữa HR* và $\widetilde{HR}$

TT	Các phép tấn công	$\rho(HR^*, \widetilde{HR})$
1	Sửa 1 bản ghi	0.486979
2	Sửa 5 bản ghi	0.458333
3	Sửa 10 bản ghi	0.515625
4	Sửa 15 bản ghi	0.486979
5	Thêm 1 bản ghi	0.492186
6	Thêm 5 bản ghi	0.518229
7	Thêm 10 bản ghi	0.476563
8	Xóa 1 bản ghi	0.473958
9	Xóa 5 bản ghi	0.481771
10	Xóa 10 bản ghi	0.481771

VI. KẾT LUẬN

Dựa theo ý tưởng mở rộng hiệu trên phần phân của Farfura [10] và bằng cách áp dụng bản đồ định vị các vị trí khả nhúng, chúng tôi đã đề xuất một phương pháp thủy vân thuận nghịch cho cơ sở dữ liệu có các thuộc tính kiểu thực. Tính đúng đắn của phương pháp đã được chứng minh bằng các phân tích lý thuyết. Phương pháp này có khả năng nhúng khá cao, bằng khoảng 70% số giá trị của thuộc tính kiểu số thực trong cơ sở dữ liệu.

Sử dụng phương pháp để xuất, kết hợp hàm băm và một hệ mật mã khóa công khai, chúng tôi đã xây dựng một lược đồ thủy vân thuận nghịch cho cơ sở dữ liệu. Lược đồ này có tính dễ vỡ, có khả năng phát hiện mọi sự thay đổi dù nhỏ trên cơ sở dữ liệu thủy vân được gửi trên mạng. Do đó người nhận có thể kiểm chứng tính toàn vẹn của cơ sở dữ liệu mình nhận được. Trong trường hợp toàn vẹn người nhận có thể khôi phục cơ sở dữ liệu gốc từ cơ sở dữ liệu thủy vân.

TÀI LIỆU THAM KHẢO

[1] A.Khan, A.Siddiqua, S.Munib, and S.A.Malik, "A Recent Survey of Reversible Watermarking Techniques", Information Sciences, 2014, pp.251-272.

[2] A.M.Alattar, "Reversible Watermarking Using the Difference Expansion of A Generalized Integer Transform", IEEE Transactions on Image Processing, 2004, Vol.18, pp.1147-1156.

[3] C.C. Lin, and F.F. Shiu, "DTC-based Reversible Data Hiding Scheme", Journal of software, 2010, Vol.5, NO.2, pp.214-224.

[4] C.C.Chang, C.C.Lin, C.S.Tseng, and W.L.Tai, "Reversible hiding in DCT-based compressed images", Information Sciences, July 2007, Vol. 177, Issue 13, pp. 2768-2786.

[5] C.C.Lee, H.C.Wu, C.S.Tsai, and Y.P.Chu, "Adaptive lossless steganographic scheme with centralized difference expansion", Patterm Recognition, 2008, pp.2097-2106.

[6] C.C.Lin, W.L.Tai, and C.C.Chang, "Multilevel reversible data hiding based on histogram modification of difference images", Pattern Recognition 41, 2008, pp.3582-3591.

[7] D.Coltuc and J-M.Chassery, "Very Fast Watermarking by Reversible Contrast Mapping", IEEE Signal processing letters, 2007, Vol. 14, No. 4, pp.255-258.

[8] Đỗ Văn Tuấn, Nguyễn Kim Sao, Nguyễn Thanh Toàn, Phạm Văn Ất (2014) "*Một sơ đồ nhúng tin thuận nghịch mới trên ảnh JPEG"*. Chuyên san Các công trình nghiên cứu, phát triển và ứng dụng Công nghệ Thông tin và Truyền thông, Tạp chí Công nghệ Thông tin và Truyền thông, tháng 12/2014, tr. 41-52.

[9] Đỗ Văn Tuấn, Trần Đăng Hiên, Phạm Đức Long, Phạm Văn Ất (2015) "*Một lược đồ thủy vân thuận nghịch mới sử dụng mở rộng hiệu đối với các véc-tơ điểm ảnh".* Chuyên san Công nghệ thông tin và Truyền thông, Tạp chí Khoa học và Kỹ thuật, Học viện Kỹ thuật Quân sự, tháng 4/2015, tr. 17-31.

[10] Farfoura, Mahmoud E., et al. "A blind reversible method for watermarking relational databases based on a time-stamping protocol." Expert Systems with Applications 39.3 (2012): 3185-3196.

[11] G. Gupta, J. Pieprzyk. "Reversible and blind database watermarking using difference expansion," Proceedings of eForensics, Jan. 2008, pp. 1–6.

[12] G.Xuan, C.Yang, Y.Zhen, Y.Q.Shi, and S.Ni, "Reversible data hiding using integer wavelet transform and companding technique", Proc. IWDW, 2004, pp.115-124.

[13] H. Khataeimaragheh, H. Rashidi (2010) "A Novel Watermarking Scheme for Detecting and Recovering Distortions in Database Tables". International Journal of Database Management Systems, Vol.2, No.3, August 2010, pp.1-11.

[14] J.Tian, "Reversible data embedding using a difference expansion", IEEE Trans. Circuits Syst. Video Technol, 2003, pp. 890–896.

[15] K.Y. Mohammad, and A.J.Ahmed, "Reversible Watermarking Using Modifiled Difference Expansion", International Journal of Computing & Information Sciences, 2006, Vol.4, No.3, pp.134-142.

[16] Li.Y, Swarup.V, Jajodia.S., (December 2003). "A robust watermarking scheme for relational data", in Proc. The 13th Workshop on Information Technology and Engineering, pp 195–200.

[17] Li.Y, Swarup.V, Jajodia.S., (October 2003). "Constructing a virtual primary key for fingerprinting relational data". ACM Workshop on Digital Rights Management, pp 133–141.

[18] M.Goljan, J.J.Fridrich, and R.Du, "Distortion-free data embedding for images", 4th Information Hiding Workshop, LNCS, 2001, Vol.2137, pp. 27– 41.

[19] M.Khodaei, and K.Faez, "Reversible Data Hiding By Using Modified Difference Expansion", 2nd International Confference on Signal Processing Systems, 2010, pp.31-34.

[20] M.Nosrati, R. Karimi, and M. Hariri, "Reversible Data Hiding, Principles, Techniques, and Recent Studies", Journal World Applied Programming, 2012, pp.349-353.

[21] R. Agrawal, J. Kiernan (2002) "Watermarking Relational Databases", Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002.

[22] W.Zhang, B.Chen, and N.Yu, "Improving Various Reversible Data Hiding Schemes Via Optimal Codes for Binary Covers", IEEE Transaction on Image Processing, June 2012, pp. 2991 – 3003.

A PUBLIC REVESIBLE WATERMAKING SCHEME FOR DATABASE USING EXTENDING REAL ATTRIBUTES

Le Quang Hoa, Do Van Tuan, Nguyen Huy Duc, Pham Van At

ABSTRACT*— Watermaking is considered a useful tool to protect the database on the unsafe exchange environment. The traditional watermaking scheme only allows get watwermark without the ability to restore the database. Therefore, the recipient does not get original data but merely an approximation. To overcome this disadvantage, recently there exists schemes that be able to restore the original data, called the reversible watermaking schemes. Difference expansion is considered an effective technique for building reversible watermaking scheme, however it has the drawback to make the data change largely. Fafoura and et.al recently proposed using difference expansion on fraction part of real attributes, thereby reducing the change and improve the quality of watermaking. However by the theoretical analysis and testing, we pointed out in a number of cases this scheme can not extract the watermark as well as restore the original data correctly. In this paper we propose a solution to overcome the drawbacks mentioned above by using a location map to distinguish cases where the method can embed by extending fraction part and the case does not allow embedding. Then based on this new information embedding method, we propose a public key fragile reversible watermaking scheme on the relational database. This scheme has the advantage that the receiver easily check the integrity of the watermaking database and get the correct restoring the original database. So the recipient does not have to use approximate database (watermaking), but can use the original database.*