LƯỢC ĐỒ THỦY VÂN DỰA VÀO BIT Ý NGHĨA NHẤT ĐỂ BẢO VỆ BẢN QUYỀN CÔNG KHAI CHO CÁC CƠ SỞ DỮ LIỆU QUAN HỆ

Lưu Thị Bích Hương

Viện Công nghệ thông tin, Trường Đại học Sư phạm Hà Nội 2, Việt Nam

ARTICLE INFORMATION

Journal: *Vinh University Journal of Science Natural Science, Engineering and Technology*
p-ISSN*: 3030-4563*
e-ISSN: *3030-4180*

Volume: 53
Issue: 3A
****Correspondence:***
luuthibichhuong@hpu2.edu.vn

Received: *12 March 2024*
Accepted: *28 May 2024*
Published: *20 September 2024*

Citation:
Luu Thi Bich Huong (2024). Watermarking scheme based on most significant bit for public copyright protection for relational databases. ***Vinh Uni. J. Sci.*** *Vol. 53 (3A), pp. 73-79 doi: 10.56824/vujs.2024a032a*

OPEN ACCESS

TÓM TẮT

Bài báo trình bày một lược đồ thủy vân dựa vào bit ý nghĩa nhất để bảo vệ bản quyền công khai cho các cơ sở dữ liệu quan hệ. Lược đồ thủy vân đề xuất có thể chứng minh một cách công khai bản quyền của dữ liệu bao nhiêu lần tùy ý. Lược đồ thủy vân bền vững trước các tấn công điển hình như: Thêm, sửa, xóa giá trị dữ liệu một cách ngẫu nhiên hoặc có lựa chọn.

Từ khóa: Bản quyền công khai; thủy vân; cơ sở dữ liệu.

1.Mở đầu

Cho đến nay, các lược đồ thủy vân dùng để bảo vệ bản quyền cho các cơ sở dữ liệu quan hệ đều phải dựa vào một khóa bí mật nào đó khi chứng minh quyền sở hữu đối với các cơ sở dữ liệu đã thủy vân [1]-[4]. Tuy nhiên, khóa bí mật này coi như đã bị lộ và vì vậy thủy vân đã nhúng sẽ dễ dàng bị phá hủy bởi những kẻ tò mò. Hơn nữa, hầu hết các lược đồ thủy vân hiện có đều làm sai lệch dữ liệu trong quy trình nhúng thủy vân. Trong đó, có một số lược đồ thay đổi các giá trị thuộc tính [2]-[3] và một số lược đồ khác lại hoán vị các giá trị dữ liệu [6]. Những sai lệch này chỉ được thực hiện trong giới hạn cho phép sao cho giá trị sử dụng của dữ liệu đối với các ứng dụng nhất định không bị ảnh hưởng và thủy vân vẫn có thể tìm lại được ngay cả khi có những tấn công làm thay đổi giá trị thuộc tính hay thêm hoặc xóa một số bộ.

Hai đặc điểm nêu trên có thể ảnh hưởng tới khả năng ứng dụng của các kỹ thuật thủy vân cơ sở dữ liệu quan hệ. Thứ nhất, cách tiếp cận dựa vào khóa thủy vân là không phù hợp cho việc chứng minh trước công chúng (ví dụ trước tòa án). Để chứng minh bản quyền của dữ liệu đáng ngờ, người chủ sở hữu phải tiết lộ khóa thủy vân của mình ra công chúng để phát hiện thủy vân. Sau khi đã sử dụng một lần, khóa này không còn bí mật nữa. Với khóa thủy vân, người sử dụng không bản quyền có thể vô hiệu hóa việc phát hiện thủy vân gốc bằng cách gỡ bỏ tất cả thủy vân gốc khỏi dữ liệu được bảo vệ hoặc thêm một thủy vân giả vào dữ liệu không được thủy vân.

Đặc điểm thứ hai là, mọi sai lệch được đưa vào trong quá trình thủy vân có thể ảnh hưởng đến giá trị sử dụng của dữ liệu. Ngay cả khi có thể ấn định một số kiểu ràng buộc nhất định để hạn chế sai sót (ví dụ như các phương sai và các giá trị trung bình của những thuộc tính được thủy vân) trước hoặc trong khi thủy vân thì cũng rất khó, thậm chí không thể đảm bảo có thể lượng hóa được tất cả các ràng buộc.

Trong bài báo [5], các tác giả đã đưa ra một lược đồ thủy vân cơ sở dữ liệu sử dụng để bảo vệ bản quyền bằng cách kiểm chứng công khai và không đưa vào bất kỳ một sự sai lệch nào đối với dữ liệu. Lược đồ thủy vân được áp dụng cho các cơ sở dữ liệu quan hệ và giả sử lược đồ quan hệ này mọi thay đổi trên bit MSB (bit ý nghĩa nhất - Most Significant Bit) của giá trị các thuộc tính là không chấp nhận được. Mặc dù các tác giả đã nhận định có thể áp dụng cho mọi kiểu dữ liệu khác nhau, tuy nhiên không làm rõ bit MSB của các kiểu dữ liệu là bit nào, điều này là khó khăn khi triển khai. Trong thuật toán 1 và thuật toán 2, các tác giả đã tạo ra quan hệ thủy vân với các kiểu dữ liệu khác nhau nhưng cách thức lấy các bit MSB của các thuộc tính là giống nhau, trong khi các cơ sở dữ liệu quan hệ luôn cập nhật dữ liệu [5]. Việc làm này sẽ dẫn đến khi cần chứng minh sẽ làm cho thủy vân không bền vững. Ý tưởng của kỹ thuật này là xuất phát từ cơ sở dữ liệu quan hệ cần bảo vệ, sinh ra một quan hệ thủy vân có cùng khóa chính, còn các giá trị của các thuộc tính khác là bit ý nghĩa nhất (MSB) của giá trị thuộc tính tương ứng của quan hệ gốc. Sau đó, hai quan hệ này sẽ được đăng ký với một tổ chức có chức năng bảo hộ bản quyền sản phẩm. Khắc phục nhược điểm đó chúng tôi đề xuất lược đồ thủy vân để bảo vệ bản quyền công khai.

Trong bài báo này, trước hết chúng tôi sẽ trình bày lược đồ thủy vân cải tiến với cách lấy các bit MSB của các kiểu dữ liệu. Quá trình chứng minh bản quyền công khai khi có tranh chấp bằng bản chứng thực được thực hiện hoàn toàn giống như lược đồ đề xuất [5]. Phân tích về độ bền vững của lược đồ cải tiến, cân đối giữa tính bền vững và các chi phí cho việc thực hiện lược đồ được thực hiện thông qua thử nghiệm và đánh giá các kết quả thu được.

2. Lược đồ

Cho r là một quan hệ thuộc lược đồ $R(P, A_1,..., A_\gamma)$ trong đó P là thuộc tính khóa chính; có γ thuộc tính có thể được chọn để thuỷ vân $A_1 ,..., A_\gamma$, ω là số các bộ trong quan hệ r. Với mỗi thuộc tính A_j (j = 1, 2, .., γ) của một bộ sẽ được biểu diễn dưới dạng nhị phân chuẩn, bit MSB của thuộc tính A_j của một bộ có thể được chọn để làm thuỷ vân. Giả sử lược đồ quan hệ ban đầu không chấp nhận sự thay đổi trên bit MSB của giá trị thuộc tính đối với giá trị sử dụng của dữ liệu."

Bảng 1: *Các ký hiệu được sử dụng trong lược đồ thủy vân*

Ký hiệu	Ý nghĩa
R	Lược đồ quan hệ
r	Quan hệ r thuộc lược đồ R
r_i	Bộ thứ i của quan hệ r
$r_i.A_j$	Giá trị của thuộc tính A_j thuộc bộ r_i
ω	Số bộ trong quan hệ r
γ	Số thuộc tính không phải khóa chính trong R
K	Khóa thủy vân

Ký hiệu	Ý nghĩa
P	Thuộc tính khóa chính
η	Tham biến sinh thuỷ vân
τ	Tham số phát hiện thủy vân
$H(K \Vert r_i.P)$	Hàm băm khóa thủy vân K cùng với giá trị thuộc tính khóa chính P của bộ t_i và || là phép ghép nối

Các thuộc tính để thủy vân của cơ sở dữ liệu quan hệ có thể nhận bất kỳ một kiểu dữ liệu nào. Lược đồ thủy vân cải tiến của chúng tôi sẽ xét đến bốn kiểu dữ liệu thường dùng đó là: kiểu số, kiểu ký tự, kiểu Boolean và kiểu datetime. Các thuộc tính của cơ sở dữ liệu đều được thể hiện dưới dạng các xâu bit trong hệ thống máy tính.

Giả sử chủ sở hữu của quan hệ r có một khóa thuỷ vân K. Khóa thủy vân phải thỏa mãn yêu cầu đủ dài và được sử dụng trong khi sinh thủy vân và phát hiện thuỷ vân trong lược đồ thủy vân. Kết cấu của thủy vân công khai r_w phụ thuộc vào khóa thủy vân K. Thủy vân r_w là một quan hệ với lược đồ $R_w(P, W_1, \ldots, W_\eta)$, trong đó P là thuộc tính khóa chính, $W_1, \ldots, W_\eta$ là các thuộc tính nhị phân. So sánh quan hệ gốc r với quan hệ thuỷ vân r_w thì hai quan hệ có cùng thuộc tính khóa chính P và cùng số bộ ω. Số các thuộc tính nhị phân trong thuỷ vân r_w là η và số bit nhị phân của r_w là ν. Để xác định số lượng các bit nhị phân ν trong quan hệ thủy vân r_w dựa vào tham biến điều khiển η với $\nu = \omega.\eta$ và $\eta \leq \gamma$. Khi đó, ta gọi η là tham biến sinh thuỷ vân.

Trong lược đồ, khóa thuỷ vân là công khai và có thể nhận bất kỳ giá trị nào (số nguyên hoặc số nhị phân) do chủ sở hữu dữ liệu lựa chọn. Không có bất cứ một ràng buộc nào về việc hình thành khóa. Để giảm bớt những nhầm lẫn không cần thiết, khóa thuỷ vân nên là duy nhất đối với chủ sở hữu của cơ sở dữ liệu quan hệ cần thủy vân. Khóa thuỷ vân tạo ra dưới dạng như sau: K = H(ID ‖ tên CSDL ‖ phiên bản ‖ ...), trong đó, ID là định danh của chủ nhân cơ sở dữ liệu, "‖" là phép ghép nối, H() là hàm băm mật mã.

Trong thuật toán sinh thuỷ vân r_w, bit MSB của các giá trị đã chọn dựa vào khóa thủy vân và khóa chính của bộ, được sử dụng để tạo thuỷ vân. Quá trình sinh thủy vân không làm thay đổi bất kỳ một giá trị thuộc tính nào của dữ liệu gốc. Việc sử dụng các bit MSB sẽ giúp ngăn chặn những tấn công thay đổi giá trị dữ liệu. Vì khóa thủy vân K, thủy vân r_w và thuật toán nhúng thủy vân đều công khai nên bất kỳ ai cũng có thể tìm được các bit MSB trong r để sinh r_w. Tuy nhiên, kẻ tấn công không thể thay đổi các bit MSB này mà không sinh ra các sai lệch không thể chấp nhận được đối với dữ liệu.

Từ các thuộc tính khác nhau, trong mỗi bộ trong quan hệ r, được chọn một cách tựa ngẫu nhiên dựa trên khóa thủy vân K và khóa chính của bộ để xây dựng thủy vân r_w. Một kẻ tấn công muốn gỡ bỏ tất cả các bit thủy vân thì sẽ phải xóa tất cả các bộ và/hoặc các thuộc tính từ dữ liệu đã thủy vân. Lược đồ thủy vân bền vững trước các tấn công gỡ bỏ hoặc xóa tất cả các bit thủy vân khi η tham số sinh thủy vân càng lớn.

Máy tính hiện nay thường biểu diễn và xử lý bốn kiểu dữ liệu cơ bản tại các địa chỉ bộ nhớ: số, datetime, ký tự và Boolean. Các kiểu dữ liệu này khi biểu diễn trong máy tính đều có dạng là một chuỗi các bit. Bit MSB của một chuỗi bit thông thường là bit trái nhất và có trọng số lớn nhất. Tuy nhiên, khi ta xét đến các kiểu dữ liệu cụ thể thì việc xác định các bit MSB lại phụ thuộc vào các kiểu. Đối với dữ liệu kiểu số (nguyên hoặc thực) có dấu thì có thể quay trái một bit để tránh bit dấu và bit trái nhất là bit MSB. Đối với dữ liệu kiểu ký tự, giả sử l là độ dài xâu ký tự, khi đó ta chọn ký tự thứ (j mod l), với j là thuộc tính được chọn để lấy bit MSB bằng cách lấy bit trái nhất của ký tự này. Giả sử có một thuộc

tính có kiểu là ký tự với giá trị là "*Trường Đại học Sư phạm Hà Nội 2 Nhân văn - Khai phóng - Hội nhập*". Đầu tiên, chọn ký tự thứ (j mod l) với j =13, l =60. Đây là kí tự "ợ" và có mã unicode là 7885. Dạng nhị phân của kí tự là 1111011001101, nên bit trái nhất có giá trị là 1. Vậy bit MSB của giá trị thuộc tính thủy vân cần tìm là 1.

Đầu vào của thuật toán phát hiện thủy vân là các tham số r', K, η, r_w, τ để phát hiện thủy vân đối với quan hệ đáng ngờ r', tham số sinh thủy vân η được sử dụng trong lược đồ thủy vân, τ là tham số phát hiện thủy vân và bằng tỷ lệ thấp nhất các bit thủy vân được phát hiện đúng. Để điều khiển độ tin cậy của thuật toán phát hiện thủy vân và độ bền vững của thủy vân đã nhúng dựa vào hai tham số η, τ, trong đó tham số τ nằm trong khoảng (0.5, 1). Do đó, khi chứng minh bản quyền công khai cho các cơ sở dữ liệu quan hệ và để tăng độ bền vững của thủy vân thì tất cả các bit MSB đã tìm được trong quan hệ thủy vân r' không cần phải trùng với các bit tương ứng trong r_w, mà chỉ cần tỷ lệ phần trăm trùng khớp lớn hơn τ là được. Tức là trong Thuật toán 2 thỏa mãn *match_count / total_count* > τ".

Thuật toán 1: Nhúng thủy vân

Input: Quan hệ r, khóa thủy vân K và tham số tạo thủy vân η ($\eta \leq \gamma$)

Output: Quan hệ thủy vân r_w

1. ***For*** i = 1 ***to*** ω ***do***
2. Xây dựng một bộ t_i trong r_w có cùng khóa chính với r_i, $t_i.P = r_i.P$
3. ***for*** k = 1 ***to*** η ***do*** // sinh giá trị cho η thuộc tính của r_w
4. j = H(K ‖ $r_i.P$) mod γ // chọn thuộc tính
5. ***if*** ($r_i.A_j$ có kiểu dữ liệu là kiểu số) ***then***
6. $q = r_i.A_j$ quay trái 1 bit // tránh bit dấu
7. $t_i.W_k$ = MSB của q
8. ***end if***
9. ***if*** ($r_i.A_j$ có kiểu dữ liệu là datetime) ***then***
10. q = year($r_i.A_j$)
11. $t_i.W_k$ = MSB của q
12. ***end if***
13. ***if*** ($r_i.A_j$ có kiểu dữ liệu là xâu ký tự) ***then***
14. q = charAt($r_i.A_j$, j mod length($r_i.A_j$)) // charAt($r_i.A_j$, p): ký tự thứ p trong $r_i.A_j$
15. $t_i.W_k$ = MSB của q
16. ***end if***
17. ***if*** ($r_i.A_j$ có kiểu dữ liệu là Boolean) ***then***
18. $t_i.W_k = r_i.A_j$
19. ***end if***
20. xóa thuộc tính thứ j trong r_i
21. ***end for***
22. ***end for***

Thuật toán 2: Phát hiện thủy vân

Input: Quan hệ nghi ngờ r', K, η ($\eta \leq \gamma$), thủy vân r_w và τ ($0.5 < \tau < 1$)

Output: {true, false}

1. match_count = 0
2. total_count = 0
3. ***for*** i = 1 ***to*** ω ***do***
4. tìm một bộ t_i trong r_w có cùng khóa chính với r'_i, $t_i.P = r'_i.P$

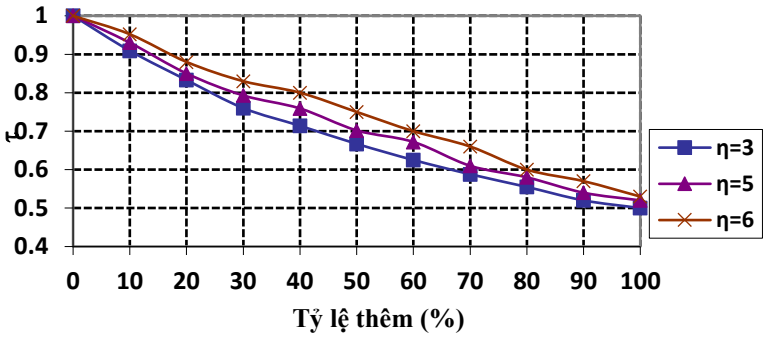
5. ***for*** k = 1 ***to*** η ***do*** // sinh giá trị cho η thuộc tính của r'_w
6. total_count = total_count + 1
7. j = H(K ‖ $r'_i.P$) mod γ
8. ***if*** ($r'_i.A_j$ có kiểu dữ liệu là kiểu số) ***then***
9. $q = r'_i.A_j$ quay trái 1 bit
10. w = MSB của q
11. ***End if***
12. ***if*** ($r'_i.A_j$ có kiểu dữ liệu là datetime) ***then***
13. q = year($r'_i.A_j$)
14. w = MSB của q
15. ***End if***
16. ***if*** ($r'_i.A_j$ có kiểu dữ liệu là xâu ký tự) ***then***
17. q = charAt($r'_i.A_j$, j mod length($r'_i.A_j$))
18. w = MSB của q
19. ***End if***
20. ***if*** ($r'_i.A_j$ có kiểu dữ liệu là Boolean) ***then***
21. $w = r'_i.A_j$
22. ***End if***
23. ***if*** ($t_i.W_k$ == w) ***then*** // so sánh r_w và r'_w
24. match_count = match_count + 1
25. ***End if***
26. xóa thuộc tính thứ j của r'_i
27. ***end for***
28. ***end for***
29. ***if*** match_count/total_count > τ ***then***
30. ***return*** true
31. ***else***
32. ***return*** false
33. ***End if***

3. Đánh giá thử nghiệm

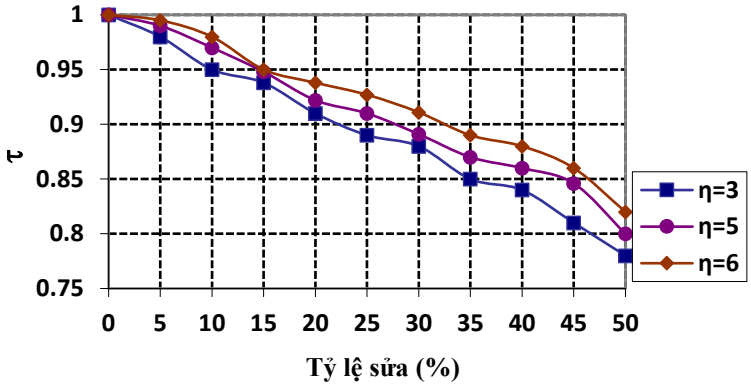
Để kiểm định tính bền vững và chi phí đối với lược đồ thủy vân cải tiến, chúng tôi đã tiến hành thử nghiệm đối với cơ sở dữ liệu quan hệ về dân số của huyện Đông Anh, thành phố Hà Nội. Dữ liệu thử nghiệm là một quan hệ gồm 2000 bộ, có một thuộc tính khóa chính là căn cước công dân và 9 thuộc tính có các kiểu dữ liệu khác nhau: họ và tên, năm sinh, giới tính, nơi sinh, địa chỉ, chủ hộ, tổng thu nhập, nghề nghiệp, trình độ chuyên môn. Thử nghiệm với các cập nhật thông thường: thêm, xóa và sửa. Các tấn công này là hoàn toàn ngẫu nhiên, giả sử không tấn công trên thuộc tính khóa chính và mỗi tấn công thử nghiệm 20 lần.

Kết quả của tấn công thêm bộ được thể hiện trong Hình 1. Nếu chọn tham số τ trong trường hợp xấu nhất là 50% và 100% bộ được thêm thì vẫn khẳng định được bản quyền của dữ liệu, do mỗi bộ được xử lý độc lập với các bộ khác dựa vào khóa chính.

Kết quả của tấn công sửa được thể hiện trong Hình 2. Theo kết quả thử nghiệm, ngay cả khi quan hệ bị sửa 50% và chọn τ lớn hơn hoặc bằng 75% thì bản quyền dữ liệu vẫn khẳng định được, do kẻ tấn công có thể không sửa vào các bit MSB.

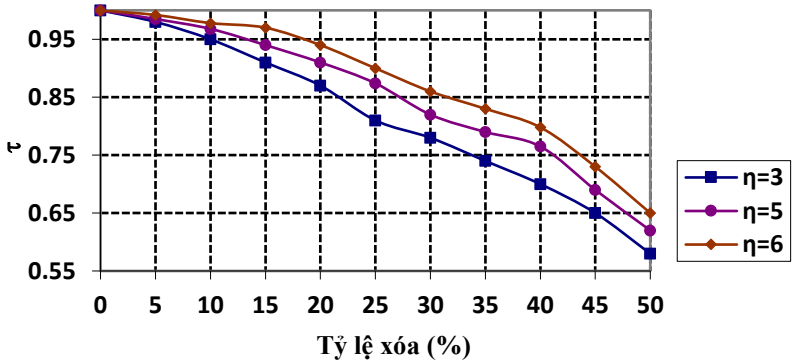


Hình 1: *Tấn công thêm bộ đối với τ*



Hình 2: *Tấn công sửa bộ đối với τ*

Kết quả của tấn công xóa được thể hiện trong Hình 3. Trong trường hợp tấn công xóa, nếu xóa 50% dữ liệu và chọn τ lớn hơn hoặc bằng 58% thì vẫn xác minh được bản quyền của dữ liệu.



Hình 3: *Tấn công xóa bộ đối với τ*

Qua thử nghiệm, để điều chỉnh giữa độ bền vững và các chi phí về thời gian cũng như bộ nhớ cho thuật toán 1 và thuật toán 2 sử dụng các tham số sinh thủy vân η và tham số phát hiện thủy vân τ. Tỷ số sinh thủy vân η được sử dụng để cân đối giữa độ bền vững và các chi phí. Tham số sinh thủy vân η càng lớn thì độ bền vững của lược đồ càng cao và các chi phí về thời gian cũng như bộ nhớ càng lớn.

4. Kết luận

Lược đồ thủy vân cải tiến không làm thay đổi bất cứ dữ liệu nào trong quá trình thủy vân, trong khi hầu hết các lược đồ thủy vân khác đều làm thay đổi dữ liệu trong quá trình thủy vân. Mặt khác lược đồ thủy vân này được áp dụng với mọi kiểu dữ liệu. Khóa thủy vân là khóa công khai nên có thể chứng minh quyền sở hữu nhiều lần, bất kỳ lúc nào và bởi bất kỳ ai. Đó là những điểm nổi bật so với những kỹ thuật thủy vân trước đó. Qua phân tích và đánh giá thử nghiệm, lược đồ thủy vân bền vững với các cập nhật thông thường.

TÀI LIỆU THAM KHẢO

[1] R. Agrawal and J. Kiernan, "Watermarking relational databases," In *Proceedings of VLDB*, pp. 155-166, 2002.

[2] Ali Al-Haj and Ashraf Odeh, "Robust and Blind Watermarking of Relational Database Systems," *Journal of Computer Science*, 4(12), pp. 1024-1029, 2008.

[3] E. Bertino, B. C. Ooi, Y. Yang, and R. Deng, "Privacy and ownership preserving of outsourced medical data," In *Proceedings of IEEE International Conference on Data Engineering*, pp. 521-532, 2005.

[4] Raju Halder, Shantanu Pal and Agostino Cortesi, "Watermarking Techniques for Relational Databases: Survey, Classification and Comparison," *Journal of Universal Computer Science*, vol. 16, no. 21, 3164-3190 (2010).

[5] Y. Li and R. H. Deng, "Publicly Verifiable Ownership Protection for Relational Databases," In *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security Table of Contents*, Taiwan, pp. 78-89, 2006.

[6] Lưu Thị Bích Hương, "Bảo vệ sự toàn vẹn của cơ sở dữ liệu quan hệ bằng kỹ thuật thủy vân," *Tạp chí khoa học Trường Đại học Vinh*, tập 50, số 1, tr. 21-29, 2021.

ABSTRACT

WATERMARKING SCHEME BASED ON MOST SIGNIFICANT BIT FOR PUBLIC COPYRIGHT PROTECTION FOR RELATIONAL DATABASES

Luu Thi Bich Huong

Institute of Information Technology, Hanoi National University of Education 2, Vietnam

Received on 12/3/2024, accepted for publication on 28/5/2024

The paper presents a watermarked scheme based on the most significant bit for public copyright protection for relational databases. A Watermark scheme could openly prove data copyright as often as desired. This watermark scheme is stable against common attacks such as adding, editing, and deleting data values randomly or selectively.

Keywords: Public copyright protection; watermark; relation database.