International Journal of Computer Science and Engineering Communications
Volume.4, Issue.4 (2016): Page.1428-1434
www.scientistlink.com/ijcsec

Authentication of Ownership Rights against Malicious Attack on non-Numeric Data Stores

R .Sindhu[1] , D. Saravana Priya[2], and Dr.M .Karthikeyan[3]
[1]PG Scholar,[2]Assistant Professor, [3]Principal,
Department of Computer Science and Engineering
P.A. College of Engineering and Technology, Coimbatore, India
[1]sindhupacet@gmail.com,[2]dspriyapacet@gmail.com,[3]karthikn.m@hotmail.com

Abstract: *Information technology is playing an increasing role in the use of relational databases. These databases are used effectively in collaborative environments for information extraction. These databases are vulnerable to security threats concerning ownership rights and data tampering. Watermarking is used to enforce ownership rights over relational data and for providing a means for tackling data tampering. When ownership rights are enforced using watermarking, the underlying data undergoes certain modifications; as a result of which, the data quality gets compromised. Reversible watermarking is employed to ensure data quality along-with data recovery. However, such techniques protect against malicious attacks and provide any mechanism to selectively watermark a particular attribute by taking into account its role in knowledge discovery. However, such techniques do not support non-numeric and shared databases. Robust and Reversible Watermarking (RRW) ensure authentication against malicious attack on non-numeric data stores.*

I. INTRODUCTION

Data Mining is the discovery of knowledge analyzing enormous set of data by extracting the meaning of the data and then predicting the future trends and also helps companies to take sound decisions, based on knowledge and information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize and summarize the relationships identified. It is the process of finding correlations or patterns among dozens of fields in large relational databases. Gathering useful information from the web has become a issue for users. The user profiles are created for user background knowledge description. User profiles represent the concept models possessed by users to gathering web information. A concept model is implicitly possessed by users and is generated from their background knowledge. Data are any facts, numbers, or text that can be processed by a computer. Organizations are accumulating vast and growing amounts of data in different formats and different databases. Information are the patterns, associations or relationships. Knowledge information can be converted into knowledge about historical patterns and future trends.

The rapid growth of online data and the widespread use of databases have created an immense need for Knowledge Discovery in Database (KDD) methodologies. Data Mining is also called as KDD. The challenge of extracting knowledge from KDD methodologies draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing. Figure 1.1 shows Data mining: confluence of multiple disciplines.

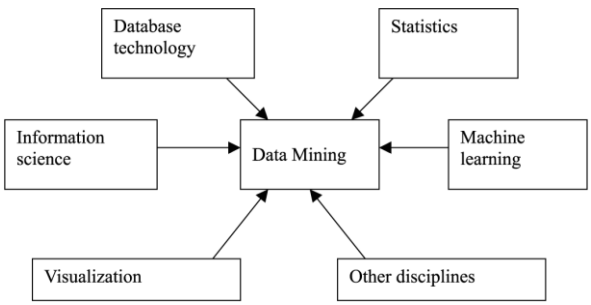


Figure 1.1 Confluence of multiple disciplines.

The Knowledge Discovery in Database is an iterative process. KDD is widely used data mining technique that includes data preparation and selection, data cleaning, incorporating prior knowledge on data sets and interpreting accurate

solutions from the observed results. The data mining along with knowledge discovery in databases are frequently treated as synonyms data mining is actually part of the knowledge discovery process. Data cleaning and data integration can be performed as a pre-processing phase to generate a data warehouse. The data selection and data transformation can also be combined. Figure 1.2 knowledge discovery process.

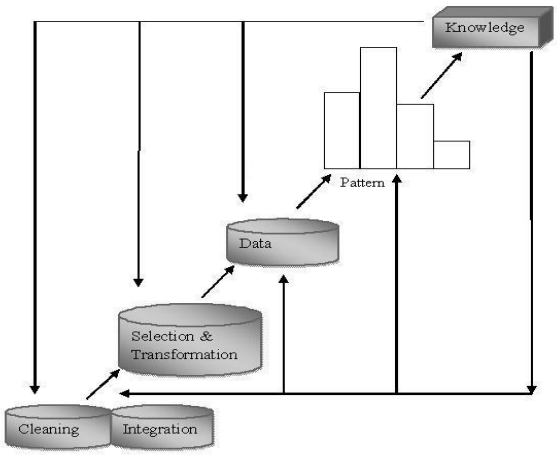


Figure 1.2 Knowledge discovery process

Data Cleaning: It is preprocessing step. In this stage, the noisy data as well as irrelevant data are removed from the database. Data Integration: At this stage, multiple data sources and heterogeneous, may be combined in a common source. Data Selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection. Data Transformation: It is also known as data consolidation. It is a phase in that selected data is transformed into forms appropriate for the mining procedure. Data Mining: It is the crucial step in that clever techniques are applied to extract patterns potentially useful. Pattern Evolution: In this step, strictly interesting patterns representing knowledge are identified based on given measures. Knowledge Representation: Discovered knowledge is visually represented to the user. The essential step uses visualization techniques to help users understand and interpret the data mining results.

The rapid growth of the Internet and related technologies has offered an unprecedented ability to access and redistribute digital contents. The enforcing data ownership is an important requirement, which requires articulated solutions, encompassing technical, organizational, and legal aspects. Watermarking techniques have emerged as an important building block that plays a crucial role in addressing the ownership problem. The techniques allow the owner of the data to embed an imperceptible watermark into the data. A watermark describes information that can be used to prove the ownership of data such as the owner, origin, or recipient of the content. Secure embedding requires that the embedded watermark must not be easily tampered with, forged, or removed from the watermarked data. Imperceptible embedding means that the presence of the watermark is unnoticeable in the data. The watermark detection is blinded, that is, it neither requires the knowledge of the original data nor the watermark. Watermarking techniques have been developed for video, images, audio, and text data, and also for software and natural language text. The problem of watermarking relational data has not been given appropriate attention. There are many application contexts for which data represent an important asset, the ownership of which must thus be carefully enforced. For example weather data, stock market data, power consumption, consumer behavior data, and medical and scientific data. Watermark embedding for relational data is made possible by the fact that real data can very often tolerate a small amount of error without any significant degradation with respect to their usability. Watermarking technique for relational data that is highly resilient compared to these techniques.

II. RELATED WORK

Secure spread spectrum watermarking should be constructed as an independent and identically distributed Gaussian random vector that is imperceptibly inserted in a spread-spectrum-like fashion into the perceptually most significant spectral components of the data[5], [19]. The insertion of a watermark under this regime makes the watermark robust to signal processing operations and common geometric transformations provided that the original image is available and that it can be successfully registered against the transformed watermarked image. Copy of text document can be made different in a nearly invisible way by repositioning or modifying the appearance of different elements of text [10], [12], [13]. A unique copy can be registered with its recipient, so subsequent unauthorized copies are retrieved can be traced back to the original owner. Architecture for distributing a large number of copies without burdening the publisher with creating and transmitting the unique documents. The architecture allows the publisher to determine the identity of a recipient who has illegally redistributed the document, without compromising the privacy of individuals. New reversible

watermarking algorithm based on the difference expansion of colored images has been developed [19], [21]. The algorithm uses spatial and spectral triplets of pixels to hide pairs of bits, allows the algorithm to hide a large amount of data.

A spatial triplet is any three pixel values selected from the same spectral component, a spectral triplet is any three pixel values selected from different spectral components. The algorithm is recursively applied to the rows and columns of the spectral components of the image and across all spectral components to maximize the hiding capacity. Reversible watermarking Prediction-Error Expansion technique to ensure authentication for digital images [15], [16], [18]. Being reversible, the algorithm enables the recovery of the original host information upon the extraction of the embedded information. Prediction-Error Expansion technique exploits the inherent correlation among the adjacent pixels in an image region using a predictor. The information bits are embedded into the prediction errors that enable a large payload keeping the distortion low. A histogram shift at the encoder enables he decoder to identify the embedded location. Reversible watermarking scheme for relational databases is proposed to achieve lossless and exact authentication of relational databases through expansion on data error histogram [11], [12]. Reversible watermarking scheme possesses the ability of perfect restoration of the original attribute data from the untampered watermarked relational databases, guaranteeing a clear and exact tampered-or-not authentication without worry about causing any permanent distortion to the database. The secret key owner possesses the capability to exactly restore the database's original state. Jeffrey Reversible watermarking enables the embedding of useful information in a host signal without any loss of host information [15]. Histogram shifting technique is used to embedding the location map that improves the distortion performance at low embedding capacities and mitigates the capacity control problem. Prediction-error expansion and histogram shifting combine to form an effective method for data embedding that improves the quality of the watermarked images, especially at moderate embedding capacities.Support Vector Based (SVB) decision support system for heart disease classification with integer-coded genetic algorithm to select critical features [20]. Simple Support Vector Machine (SSVM) algorithm has been used to determine the support vectors in a fast, iterative manner. For selecting the important and relevant features and discarding the irrelevant and redundant ones, integer-coded genetic algorithm is used which also maximizes SVM's classification accuracy. As a two class problem, disease or no disease, the proposed method gives an accuracy of 90.57% which shows an improvement over the existing methods. Text format based relational database watermarking method for non-numeric multi words data is proposed [1], [2]. A mark is embedded by horizontally shifting the location of a word within selected attribute of selected tuples; a word is displaced right or left unmoved depending on watermark bit. The location where the mark to be inserted is determined by the distance between two successive words within an attribute. The method is effective as it is robust against different forms of malicious attacks and it is blind as it does not require the original database in order to extract the embedded watermark.

III. DIGITAL WATERMARKING

In the digital world, data is excessively generated due to increasing use of the Internet and cloud computing. Data is stored in different digital formats such as images, audio, video, natural language texts and relational data. Relational data in particular is shared extensively by the owners with research communities and in virtual data storage locations in the cloud. The purpose is to work in a collaborative environment and make data openly available for knowledge extraction and decision making. A large multinational retail corporation that has made its sales database available openly over the internet for the purposes of identifying market trends through data mining. Openly available datasets make attractive targets for attacks. Watermarking techniques is used to ensure security in terms of ownership protection and tamper proofing for a wide variety of data formats. Reversible watermarking techniques can ensure data recovery along with ownership protection. Data modifications are allowed to such extent that the quality of the data before embedding watermark information and after extracting is acceptable for knowledge extraction process.

IV. ROBUST AND REVERSIBLE WATERMARKING

A. ARCHITECTURE

Robust and Reversible Watermarking (RRW) for relational databases improves data recovery ratio. RRW includes watermark preprocessing, watermark encoding, watermark decoding and data recovery. Figure 3.1 shows main architecture of RRW. The watermark preprocessing phase computes different parameters for calculation of an optimal watermark. These parameters are used for watermark encoding and decoding. The watermark encoding phase is to embed watermark information and it does not affect the data quality. During watermark embedding, data gets modified according to the available bandwidth of watermark information. The bandwidth of the watermark should be sufficiently large to ensure robustness but not so large that it destroys the data quality. The data owner decides the amount of data modification and the quality is not compromised for a particular database application. Numerical features can be taken under consideration from any dataset and a suitable feature is determined to embed watermark on the basis of mutual information. After watermarking the data is released to the intended recipients over a communication channel that is assumed to be insecure and termed as the attacker channel. The data may undergo several malicious attacks in the attacker channel. The efficiency

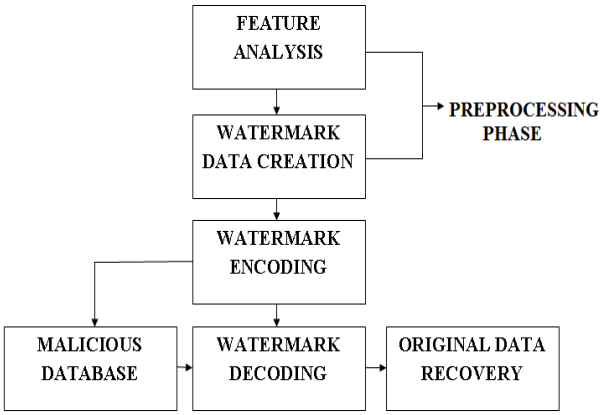
and effectiveness of RRW is calculated through robustness analysis determined by its response to subset insertion, alteration and deletion attacks .The Watermark decoding phase recovers watermark information effectively for detection of the embedded watermark. Data recovery phase mainly comprises the important task of successful recovery of the original data.

B. WATERMARK PREPROCESSING PHASE

Feature Analysis

Features Mutual Information (MI) are calculated according to their importance in information extraction, subject to their mutual dependence on other features. Mutual Information (MI) is an important statistical measure for computation of mutual dependence of two random variables. Mutual information of every feature with all other features is calculated by using Equation,

$$MI(P,S) = \sum_{p}\sum_{s} P_{PS}(p,s) log \frac{P_{PS}(p,s)}{P_P(p)P_S(s)}$$



.Figure 3.1 Architecture of RRW

Where, MI (P,S) measures the degree of correlation of features by measuring the marginal probability distributions as P_P(p), P_S(s) and the joint probability distribution P_{PS}(p,s). Then MI of one feature with all other features is computed using the relation,

$$MI_{fi} = (MI_{fij})$$

Where i, j =1, 2 . . . ,f_t with i ≠j, and f_t is the total number of features. The value of MI of each feature is also used to rank the features. The attacker will try and predict the feature with the lowest MI in an attempt to guess the feature is watermarked. To deceive the attacker for the particular scenario, a secret threshold can be used for selecting the feature for watermark embedding. The data owner can define a secret threshold based on MI of all the features in the database. The feature having MI lower than that threshold can be selected for watermarking. The attacker will not attack the features having large MI. The attacker will be forced to attack the feature with lower MI without concrete of the features that have been watermarked.

Watermark Creation

For the creation of optimal watermark information, that needs to be embedded in the original data Genetic Algorithm (GA). GA is a population-based computational model, basically inspired from genetic evolution. The GA preserves essential information through the application of basic genetic operations to these chromosomes that include selection, crossover, mutation and replacement. The GA evaluates the quality of each candidate chromosome by employing a fitness function. The evolutionary mechanism of the GA continues through a number of generations, until some termination criteria. Steps involved in GA for creating a optimal watermark information:

1) Initial random population of binary strings called chromosomes is generated. Gene values of each chromosome represent l-bit watermark string.

2) Fitness of each chromosome is evaluated by employing a constrained optimized fitness function.
3) Tournament selection mechanism is applied to get the most appropriate individuals as parent chromosomes.
4) Genetic operations of crossover and mutation are performed on parent chromosomes to create offspring's. A single point crossover operator is applied to evolve high quality individuals, inheriting parental characteristics, by exchanging information between two or more chromosomes. A uniform mutation operator is applied to bring diversity in population through small random changes in gene values of binary chromosomes.
5) Elitism strategy is applied to hire two individuals with best fitness value; as elites to the next generation without genetic changes.
6) Remaining population of the next generation is created by replacing less fit individuals of the previous generation with the mostly fit newly created off-springs.
7) Steps 2 to 6 are repeated until MI_O and MI_W reach approximately equal values for a certain number of generations.
8) Optimal watermark information string and best fitness value (β) is returned after the fulfillment of the termination criteria.

C. WATERMARK ENCODING PHASE

GA is used to create optimal watermark information that includes optimal chromosomal string and β value. β is a parameter that is computed using GA and represents a tolerable amount of change to embed in the feature values. Once the optimum value of β for each candidate feature is found, is saved for use during watermark encoding and decoding. The value β is added into every tuple of all features when a given bit is 0 and subtracted from value of all features when a given bit is 1.

$$\eta_r = D_r * \zeta \quad (3)$$

$$D_{Wr} = D_r + \beta \quad (4)$$

$$D_{Wr} = D_r - \beta \quad (5)$$

Algorithm 1.Watermark Encoding

```
Input: D, w, β
   for w = 1 to l do
//loop will iterate for all watermark bits
   for t = 1 to E do
         //loop will iterate for all tuples
     if w = = 0 then
         //when watermark bit is 0
         //calculate changes using Equation 3
         //watermark data using Equation 4
     end if
     if w = = 1 then
         //when watermark bit is 1
         //calculate changes using Equation 3
         //watermark data using Equation 5
     end if
    end for
  end for
  return Dw, Δ
Output: Dw, Δ
```

The above step is repeated until all tuples of the dataset have been watermarked.

D. WATERMARK DECODING PHASE

In the watermark decoding process, features that have been watermarked is identified. The process of optimization through GA is not required in this phase. We use a watermark decoder ζ, which calculates the amount of change in the value of feature that does not affect data quality. The watermark decoder decodes the watermark by working with one bit at a time.

$$\eta_{dr} = D'_w * \zeta \quad (6)$$

$$\eta_{\Delta r} = \eta_{dr} - \eta_r \quad (7)$$

Algorithm 2.Watermark Decoding

Input: D_w or D'_w , l
for t = 1 to E **do**
// loop will iterate for all tuples
for w = 1 to l **do**
//loop will iterate for all watermark bits
$\eta_{dr} = D'_w * \zeta$
$\eta_{\Delta r} = \eta_{dr} - \eta_r$
if $\eta_{\Delta r}$ <= 0 **then**
detected watermark bit is 1
else if $\eta_{\Delta r} > 0$ and $\eta_{\Delta r} <= 1$ **then**
detected watermark bit is 0
end if
end for
end for
return W_D
Output: W_D

E. DATA RECOVERY PHASE

After detecting the watermark string, some post processing steps are carried out for error correction and data recovery. The optimized value of β computed through the GA is used for regeneration of original data.

$$D_r = D'w + \beta \quad (8)$$

$$D_r = D'w - \beta \quad (9)$$

Algorithm 3.Data Recovery

Input: D_W or D'_W ,b
for t = 1 to E do
//loop will iterate for all tuples
for w = l to 1 **do**
//loop will iterate for all watermark bits
if W(r,b) = = 1 **then**
//0 or 1 watermark bit is detected from every tuple r data is recovered using Equation 8
else
data is recovered using Equation 9
end if
end for
end for
return D_r
Output: D_r

V. RESULT AND ANALYSIS

Experiments are conducted on Intel Core i3 with CPU of 2.40 GHz and RAM of 2 GB. For brevity, heart disease medical dataset containing more than 300 tuples is selected. RRW was evaluated for: (1) investigating effect on the data quality of the underlying data; (2) robustness against malicious attacks; and (3) restoration of the original data. The data recovery, watermark detection accuracy and effect of RRW on data quality are evaluated using the case study of a heart disease medical dataset.

Robustness of RRW is demonstrated through an extensive attack analysis. Our results have shown 100 percent accuracy in both watermark detection and data recovery. Results are compared with effect on the statistical measures of the RRW technique before and after watermarking. RRW is also compared with DEW, GADEW and PEEW techniques for watermark detection accuracy with subset insertion, subset alteration and subset deletion attacks. In all these scenarios, RRW technique has shown better results.

VI. CONCLUSION

Irreversible watermarking techniques make changes in the data to such an extent that data quality gets compromised. However, these techniques are robust against malicious attacks—particularly those techniques that target some selected tuples for watermarking. In this paper, a novel robust and reversible technique for watermarking numerical and non-numerical data of relational databases is presented. The main contribution of this work is that it allows recovery of a large portion of the data even after being subjected to malicious attacks. RRW is also evaluated through attack analysis where the watermark is detected with maximum decoding accuracy in different scenarios. A number of experiments have been conducted with different number of tuples attacked. One of our future concerns is to watermark shared databases in distributed environments where different members share their data in various proportions.

REFERENCES

[1] Y.-C. Liu, Y.-T. Ma, H.-S. Zhang, D.-Y. Li, and G.-S. Chen, "A method for trust management in cloud computing: Data coloring by cloud watermarking," Int. J. Autom. Comput., vol. 8, no. 3, pp. 280–285, 2011.

[2] (2013, Apr. 11). Identity theft watch. [Online]. Available: http:// scambook.com/blog/2013/ 04/identity-theft-watch-customerpasswords-stolen-from-walmart-vudu-video-service/

[3] (2013, Feb. 26). Securing outsourced consumer data. [Online]. Available:http://databreaches.net/securing-outsourced-consumer-data/

[4] (2012, Jun. 3). As patients' records go digital, theft and hacking problems grow. [Online]. Available: http://kaiserhealthnews. org/Stories/2012/June/04/ electronic-health-records-theft-hacking.aspx

[5] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," IEEE Trans. Image Process., vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[6] I. Cox, M. Miller, J. Bloom, and M. Miller, Digital Watermarking. Burlington, MA, USA: Morgan Kaufmann, 2001.

[7] P. W. Wong, "A public key watermark for image verification and authentication," in Proc. IEEE Int. Conf. Image Process., 1998, vol. 1, pp. 455–459.

[8] P. W. Wong and N. Memon, "Secret and public key image watermarking schemes for image authentication and ownership verification," IEEE Trans. Image Process., vol. 10, no. 10, pp. 1593– 1601, Oct. 2001.

[9] F. A. Petitcolas, "Watermarking schemes evaluation," IEEE Signal Process. Mag., vol. 17, no. 5, pp. 58–64, Sep. 2000.

[10] J. T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright protection for the electronic distribution of text documents," Proc. IEEE, vol. 87, no. 7, pp. 1181–1196, Jul. 1999.

[11] R. Agrawal and J. Kiernan, "Watermarking relational databases," in Proc. 28th Int. Conf. Very Large Data Bases, 2002, pp. 155–166.

[12] R. Sion, M. Atallah, and S. Prabhakar, "Rights protection for categorical data," IEEE Trans. Knowl. Data Eng., vol. 17, no. 7, pp. 912– 926, Jul. 2005.

[13] S. Subramanya and B. K. Yi, "Digital rights management," IEEE Potentials, vol. 25, no. 2, pp. 31–34, Mar.-Apr. 2006.

[14] P. E. Gill, W. Murray, and M. A. Saunders, "Snopt: An sqp algorithm for large-scale constrained optimization," SIAM Rev., vol. 47, no. 1, pp. 99–131, 2005.

[15] K. E. Parsopoulos and M. N. Vrahatis, "Particle swarm optimization method for constrained optimization problems," Intel. Technol.–Theory Appl. New Trends Intell. Technol., vol. 76, pp. 214– 220, 2002.

[16] R. Hassan, B. Cohanim, O. De Weck, and G. Venter, "A comparison of particle swarm optimization and the genetic algorithm," in Proc. 46th AIAA/ASME/ASCE/AHS/ASC Struct., Struct. Dyn. Mater. Conf., 2005, pp. 1–13.

[17] Y.-R. Wang, W.-H. Lin, and L. Yang, "An intelligent watermarking method based on particle swarm optimization," Expert Syst. Appl., vol. 38, no. 7, pp. 8024–8029, 2011.

[18] M. Kamran and M. Farooq, "An information-preserving watermarking scheme for right protection of EMR systems," IEEE Trans. Knowl. Data Eng., vol. 24, no. 11, pp. 1950–1962, Nov. 2012.

[19] T. M. Cover, J. A. Thomas, and J. Kieffer, "Elements of information theory," SIAM Rev., vol. 36, no. 3, pp. 509–510, 1994.

[20] T. M. Cover and J. A. Thomas, Elements of Information Theory. New York, NY, USA: Wiley-Interscience, 2012.

[21] Y. Zhang, B. Yang, and X.-M. Niu, "Reversible watermarking for relational database authentication," J. Comput., vol. 17, no. 2, pp. 59–66, 2006.

[22] G. Gupta and J. Pieprzyk, "Reversible and blind database watermarking using difference expansion," in Proc. 1st Int. Conf. Forensic Appl. Tech. Telecommun., Inf., Multimedia Workshop, 2008, p. 24.