Right-Protected Data Publishing with Provable Distance-based Mining: A Survey

S. Mythili[1] Dr.V.Thiagarasu[2]
[1]M.Phil Research Scholar, PG & Research [2]Associate Professor
[1,2]Department of Computer Science
[1,2]Gobi Arts & Science College (Autonomous), Gobichettipalayam - 638 453

Abstract— Data exchange and data publishing are becoming an essential part of business and academic practices and data owners need to maintain the principal rights over the concern datasets that they share. This survey reviews the right-protection mechanism that can provide detectable evidence for the legal ownership of a shared dataset, without compromising its usability under wide range of machine learning, mining and search operations. The watermarking preservation algorithms protect important properties of the dataset which are essential for mining operations and so guarantee both right protection and utility preservation. It proves fundamental lower and upper bounds on the distance between objects. In particular, it establishes a restricted isometric property, i.e., tight bounds on the expansion of original distances. The review of various methods is discussed to assess the quality and reliability of right-protection mechanisms. From the survey it is observed that the right-protection mechanism based on watermarking technique has good efficiency and better capability for upgrading the best approach and helps to overcome the problems by analyzing and focusing the components of watermarking preservation algorithms. This analysis used to design fast algorithms for NN (Nearest Neighbors)-preserving watermarking that drastically prunes the vast search space. Then the performances of various existing watermarking algorithms are discussed and also the limitations of object distance validation are discussed.

Key words: Data Publishing, Right-protection, Watermarking techniques, NN (Nearest Neighbors) – preserving

I. INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis and it uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events[5]. The key properties of data mining are: (1) Automatic discovery of patterns (2) Prediction of likely outcomes (3) Creation of actionable information (4) Focus on large data sets and databases[6]. The knowledge driven data mining systems cannot be developed and designed until the owner of the data is willing to outsource the data with corporations or data mining experts[7]. In the emerging field of outsourced datasets with the intended recipients, protecting ownership of the data is becoming a challenge in itself. The commonly used mechanism to enforce and proves ownership for the digital data in different formats is watermarking[19].

In the recent years, copyright protection of digital content became a severe problem due to hasty development in technology[20]. A "Watermark" is a signal that is firmly, imperceptibly and robustly embedded into original content such as an image, video, or audio signal producing a watermarked signal[21]. The watermark describes information that can be used for proof of ownership or tamper proofing[22]. To discover right-protect a dataset, but at the same time to guarantee preservation of the outcome of important distance-based mining operations, the approach provides two variants: one that preserves Nearest-Neighbors (NN) and another that preserves the Minimum Spanning Tree (MST). To guarantee this, the study of critical watermark intensity is used to protect the dataset, as well as ensure that important parts of the object distance graph are not distorted. The essential part is to discover the maximum watermark intensity for right protection and provides assurance of better detect ability and hence better security for the right protection scheme[23].

The watermark detection process aims at discovering the existence of a particular watermark in a watermarked dataset[24]. This involves measuring the correlation between a tested watermark technique and the watermarked dataset. The higher correlation between the two, the higher probability that embedded watermark was the one tested. Because the watermark is embedded in all objects of a dataset, one option is to measure the correlation between watermark and average of the magnitudes of Fourier descriptors across all objects of the dataset. However, directly measuring the correlation may not be very effective under multiplicative embedding.

A watermark embedding method is referred to as spread-spectrum if the marked signal is obtained by an additive modification[25]. Spread-spectrum watermarks are known to be modestly robust, but also to have a low information capacity due to host interference[26]. The method is said to be of quantization type if the marked signal is obtained by quantization. Quantization watermarks suffer from low robustness, but have a high information capacity due to rejection of host interference[27]. This watermarking method is referred to as amplitude modulation if the marked signal is embedded by additive modification which is similar to spread spectrum method, but is particularly embedded in the spatial domain.

II. REVIEW ON WATERMARKING PRESERVATION ALGORITHMS

1) Vidhi Khanduja, Shampa Chakraverty and Om Prakash Verma [2015][1] proposed a three-level security strategy and a robust watermarking scheme for relational databases containing categorical data that resolves ownership issues to deter piracy. Each watermark bit is embedded multiple times into different partitions and this makes the scheme highly robust against various attacks. Since technique is distortion

less, it is suitable for any data type attribute such as numeric, non-numeric etc.

2) Baisa L Gunjal and Suresh N Mali [2015][2] proposed Multi-objective Evolutionary Optimizer (MEO) on basis of highly secured and strongly robust image watermarking technique using Discrete Wavelet Transform (DWT) and Singular Value Decomposition (SVD). Existing techniques are relatively slow and the proposed DWT decomposition with Haar wavelet gives better performance with reduced computation time. The proposed technique achieved normalized correlation for all cover images indicating exact recovery of watermark. This technique is flexible and can be easily extended for color image watermarking.
3) Vinita Gupta and Atul Barve [2014][3] presented the various factors used in watermarking, properties and application area where water making technique need to be used. The classified watermarking algorithms based on the transform domain are discussed. The watermarks are embedded using the transform domain algorithms. The detailed study on watermarking properties, applications and techniques are given and a comparative study on various algorithms is discussed.
4) Khalid A. Darabkh [2014][4] proposed efficient audio watermarking embedding and extracting techniques, which mainly use Discrete Wavelet Transform (DWT) and Singular Value Decomposition (SVD). The aim of this method is to improve the imperceptibility and robustness obtained by currently proposed audio watermarking algorithms. Furthermore, many experiments were conducted to study the affect of employing multiple levels of DWT (Discrete Wavelet Transform) and different watermark intensities on the imperceptibility and robustness utilizing the proposed matrix formation.
5) Jithin VM, K K gupta [2013][8] proposed QR code on watermark image and this makes watermarked image more robust. If the embedding algorithm is known to unauthorized person then by using the QR code scanner software, the watermarking key can be easily extracted.
6) Michail Vlachos, Aleksander Wieczorek and Johannes Schneider[2012][9] proposed Hierarchical Clustering Preservation Algorithms for right protection of dataset. Hierarchical clustering (HC) builds a nested hierarchy of groups of objects according to a given distance functions. This nested hierarchy is called a dendrogram. Two algorithms have been taken for the hierarchical clustering preservation: (1) The Single-linkage hierarchical clustering algorithm, (2) The Complete-linkage hierarchical clustering algorithm. The algorithms are designed to find the maximum embedding power that guarantees preservation of hierarchical clustering operations on the modified dataset. The fast variants that put forward can reduce the search space by more than 3000 times compared to the exhaustive algorithms with no sacrifice in accuracy.
7) Qing Liu, Jun Ying [2012][10] proposed the DWT (Discrete wavelet transform) and applying DWT the original image is transformed up-to the 3-layers or 3 times, so that image is divided into the different sub band and watermarked image is embedded into the intermediate frequency sub band. Spread spectrum technology is also used and blind watermarking technique is used to extract the watermark. Spread spectrum technology provides secure communications because signal is "hidden" like noise but it increases bandwidth of signal and increases the complexity and also used blind detection technique to extract the watermark which is used.
8) Zhaoshan Wang, Shanxiang Lv and Yan Shna [2012][11] proposed a digital image watermarking algorithm based on chaos and Fresnel transform. The original image is transformed by using the concept of Fresnel diffraction plane by distance parameter and watermark image is embedded after scrambled by chaotic sequence. The watermark image can be retrieved without original image and there are little changes on the original image after embedding. Chaotic scrambling can encrypt watermark information.
9) Nan Lin, Jianjing Shen, Xiaofeng Guo and Jun Zhou [2011][12] proposed a novel blind watermarking technique based on QR decomposition in still images. The method is implemented in wavelet domain and its robustness has been evaluated against some image processing attacks. The results have been compared with two traditional methods i.e., SVD() and DCT() and shown while the proposed scheme has low computational complexity and it has better robustness against some image processing attacks in comparison with SVD and DCT methods.
10) Claudio Lucchese, Michail Vlachos, Deepak Rajan and Philip S. Yu[2010][13] proposed Neighbor-Preserving watermarking through Fast search algorithms on trajectory dataset and presents a technique of convincingly claiming ownership rights over a trajectory dataset. The presented methodology distorts imperceptibly a collection of sequences and effectively embedding a secret key. This analysis provides ownership assurances on trajectory datasets using watermarking principles. The effectiveness of the Neighbor-Preserving algorithms is quantified for determining the watermark embedding power. The speed of execution is strongly dependent on the dataset size; the larger cardinality of objects in the dataset, the larger improvement gained by the Fast Search algorithm.
11) Hemin Golpira et al. [2009][14] reported reversible blind watermarking and this approach is processed during embedding process, by applying Integer Wavelet Transform (IDWT) image is decomposed into four sub-bands. According to the capacity required for the watermark data, watermark is embedded by selecting two points called thresholds. To get watermarked image Inverse Integer Wavelet Transform (IIDWT) is applied. In the extraction process, all of these phases are performed in reverse order to extort watermark as well as host image.
12) Rodriquez et al.[2007][15] proposed a method to search a appropriate pixel to embed information using the spiral scan that starts from the centroid of cover image. Then by obtaining the block with its center at the position of selected pixel, it checks the value of bit to embed. While in extraction process, the position of marked pixel is obtained by spiral scan starting from

centroid of the cover image. By checking the luminance value of the central pixel with the gray-scale level mean of the block, the embedded bit is identified.

13) Giakoumaki et al. [2006][16] presented a multiple watermarking method using wavelet-based scheme. The method provides solution to the number of medical data management and distribution issues such as data confidentiality, archiving and retrieval, and record integrity. In this approach up to 4 levels DWT (Digital Watermarking Technique) is performed on medical image. The algorithm embeds multiple watermarks in different level. The method embeds caption watermark holding patient's personal information in second decomposition level. Moreover, a fragile watermark is embedded in forth-level decomposition. Extraction process is the reverse of embedding process and experimentation is done on ultrasounds medical images.
14) Rakesh Agrawal and Jerry Kiernan et al [2006][17] proposed an algorithm for database watermarking based on primary key and private(secrete) key. The process of inserting a single bit watermark into the numeric field of database and then detecting it with the help of detection algorithm is done. A watermark can be applied to any database relation with attributes and which having changes in a few of the values and do not affect the applications. An effective watermarking technique geared for relational data. This technique ensures that some bit positions of some of the attributes of some of the tuples contain specific values. The tuples attribute within a tuple, bit positions in an attribute and specific bit values are all algorithmically determined under the control of a private key known only to the owner of the data.
15) Francesc Sebe, Josep Domingo-Ferrer and Agusti Solanas[2005][18] proposed Mark Embedding and Mark Recovery algorithms on numerical datasets. The problem of inserting a watermark into a numerical dataset is handled while preserving the average and the variance of the original data. It is clearly observed that the distortion produced by the noise necessary to remove the watermark is much greater than the one produced by the insertion of the watermark on the original data. An analytical expression of the information loss (distortion) caused by watermark embedding on the original data has been given.

III. COMPARISON OF WATERMARKING PRESERVATION ALGORITHMS

This survey reviews the analysis on various existing algorithms proposed by different authors for right-protection on data publishing. The problem of the review is a spread-spectrum approach and embeds the watermark across multiple frequencies of each object and across multiple objects of the dataset. The spread-spectrum approach provides two variants: one that preserves Nearest-Neighbors (NN) and another that preserves the Minimum Spanning Tree (MST). Therefore, the output of any algorithm based on these two properties will be preserved after right protection. To guarantee this, the study should be on critical watermark intensity to protect the dataset, as well as ensure important parts of the object distance graph are not distorted. An essential part is to discover the maximum watermark intensity for right protection. This provides assurances of better detectability and hence better security for the right protection scheme. As such, it renders the removal of the watermark particularly difficult without substantially compromising the data utility. The data locations are altered before applying the watermarking. The robustness of the watermark embedding depends on the choice of coefficients. The watermark is embedded in the coefficients that exhibit and on average over the dataset, the largest Fourier magnitudes. This makes the removal of the watermark difficult. The process of analyzing should be on (Euclidean) distances between the objects that are distorted as a function of the watermark embedding strength. This leads to design fast variants of algorithms that still guarantee preservation of the NN and the MST, but operate significantly faster than the exhaustive algorithms. However, during the data distribution, all kind of receiver users receive the same data. The proposed system can overcome this problem by having multiple watermarking on same dataset and the Fast algorithms for NN-Preserving and MST-Preserving can be improved for right-protection efficiency.

Author (Year)	**Algorithms**	**Accuracy**	**Type of dataset**	**Nature**
Francesc Sebe (2005)	Mark Embedding, Mark Recovery	Medium	Numerical Datasets	Inserting a watermark into a numerical dataset is handled while preserving the average and the variance of original data. This method is robust against additive noise attacks.
Rakesh Agrawal, Jerry Kiernan et al (2006)	Detection algorithm	Medium	Database consists of both numeric and non-numeric datasets	Inserting a single bit watermark into the numeric and non-numeric fields of database and then detecting it. The tuples, attributes within a tuple, bit positions in an attribute and specific bit values are all algorithmically determined under the control of a private key known only to the owner of the data.
Ashu Gupta (2012)	Distortion based watermarking, LSB (Least Significant Bits) based watermarking	Medium	Numerical Database	There is no huge change between the original database and the watermarked database. It is also robust against subset addition, deletion and alteration attacks.

Michail Vlachos (2012)	Hierarchical Clustering (HC) Preservation Algorithms	High	Dataset of both numeric and non-numeric	The algorithms are designed to find the maximum embedding power that guarantees preservation of hierarchical clustering operations on the modified dataset. Reduce the search space by more than 3000 times compared to the exhaustive algorithms with no sacrifice in accuracy.
Claudio Lucchese (2010)	Fast Search Algorithms (NP-Neighbor Preserving)	High	Trajectory Dataset	Neighbor Preserving (NP) watermarking is used. The speed of execution is strongly dependent on dataset size; the larger cardinality of objects in the dataset, the larger improvement gained by the Fast Search algorithm.

Table 1: Comparison of Existing Algorithms

Taking into consideration of existing analysis, watermarking methods are useful and efficient one for right protection of dataset. Moreover, the experimental results indicate that NN-Preserving algorithms can easily find more relevant object distances of original dataset comparing to other methods. Fig.3.1. describes the accuracy analysis of existing algorithms and gives the year wise report of improvement on different watermarking preservation algorithms.

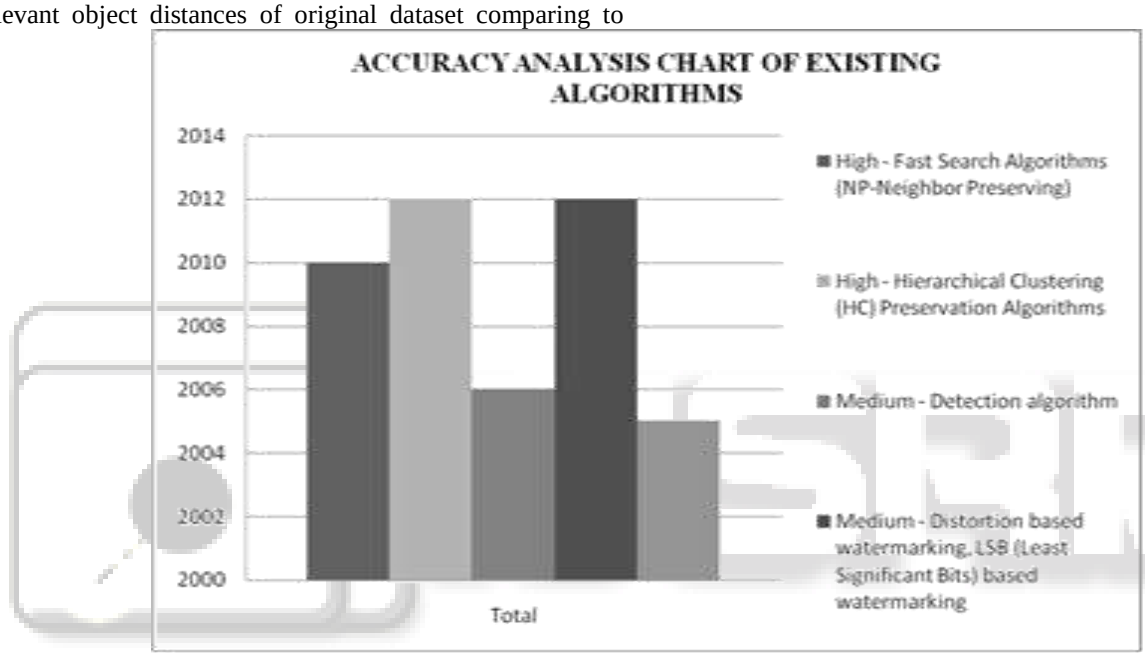

Fig. 3.1: Accuracy analysis of existing algorithms

IV. Conclusion

The copyright protection of digital content became a serious problem due to rapid development in technology. Watermarking is one of the choices to copyright-protection problem. In this survey, various watermarking methodologies and algorithms by different authors have been reviewed and paying attention to get efficient right-protection of dataset. Watermarking is applied in both image and numeric data set. Many issues in watermarking preservation algorithm design remain open and should get attention in future work. From these open issues, we select the analysis of the NN-Preservation algorithm as one of the most important ones, since fast algorithms for NN-Preserving on watermarking technique improves the speed and quality protection on ownership of dataset. The Nearest Neighbors (NN) preserves important property of each object of the original dataset. This leads to preservation of any mining operation that depends on the ordering of distances between objects, such as NN-search and classification, as well as many visualization techniques. The experimental results are conducted for selecting the best algorithm to achieve efficient right-protection mechanism. The best approach is identified using the experimental results obtained. Although the performance of the best approach is better than other algorithm, it has some problems. In order to improve the performance of the best approach and overcome the problems as aforementioned the Fast NN (Nearest Neighbors)-Preserving algorithm will be proposed in future.

References

[1] Vidhi Khanduja, Shampa Chakraverty and Om Prakash Verma, "Watermarking Categorical Data : Algorithm and Robustness Analysis", Defence Science Journal, pp. 226-232, DOI : 10.14429/dsj.65.8444, Vol. 65, No. 3, May 2015.

[2] Baisa L Gunjal and Suresh N Mali, "MEO based secured, robust, high capacity and perceptual quality image watermarking in DWT-SVD domain", DOI 10.1186/s40064-015-0904-z, SpringerPlus, 4:126-2015.

[3] Vinita Gupta and Atul Barve, "A Review on Image Watermarking and Its Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Vol. 4, Issue-1, Jan. 2014

[4] Khalid A. Darabkh, "Imperceptible and Robust DWT-SVD-Based Digital Audio Watermarking Algorithm", Journal of Software Engineering and Applications, Scientific Research, 7, 859-871, 2014.
[5] Lalit Kumar Saini and Vishal Shrivastava,"A Survey of Digital Watermarking Techniques and its Applications", International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 3, May-Jun 2014.
[6] R. Sion, M. J. Atallah, and S. Prabhakar, "Rights protection for discrete numeric streams," IEEE Trans. Knowl. Data Eng., vol. 18, no. 5, pp. 699–714, May 2006.
[7] M. Vlachos, C. Lucchese, D. Rajan, and P. S. Yu, "Ownership protection of shape datasets with geodesic distance preservation," in Proc. 11th Int. Conf. EDBT, Nantes, France, 2008, pp. 276–286.
[8] Jithin VM and K K gupta, "Robust invisible QR code image watermarking in DWT domain", 2013.
[9] Michail Vlachos, Aleksander Wieczorek and Johannes Schneider, "Right-Protected Data Publishing with Hierarchical Clustering Preservation", CIKM'12, Maui, HI, USA, ACM 978-1-4503-1156-4/12, 2012.
[10] Qing Liu, Jun Ying (2012),"Grayscale Image Digital Watermarking Technology Based on Wavelet Analysis", 2012.
[11] Zhaoshan Wang, Shanxiang Lv and Yan Shna (2012), "A Digital Image Watermarking Algorithm Based on Chaos and Fresnel Transform", 2012.
[12] Nan Lin, Jianjing Shen, Xiaofeng Guo and Jun Zhou, "A robust image watermarking based on DWT-QR decomposition," Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on , vol., no., pp.684,688, 27-29 May 2011.
[13] C. Lucchese, M. Vlachos, D. Rajan, and Philip S. Yu, "Rights protection of trajectory datasets with nearest-neighbor preservation," The International Journal on Very Large Databases (VLDB), vol. 19, no. 4, pp. 531–556, 2010.
[14] Hemin Golpira and Habibollah Danyali, "Reversible Blind Watermarking for Medical Images Based on Wavelet Histogram Shifting", IEEE, 2009.
[15] C.R. Rodriguez, F. Uribe Claudia, T. Blas Gershom De J, "Data Hiding Scheme for Medical Images", IEEE 17th International Conference on Electronics, communications and computers, 2007.
[16] Giakoumaki, Sotiris Pavlopoulos, and Dimitris Koutsouris, (Oct. 2006) "Multiple Image Watermarking Applied to Health Information Management", IEEE Trans. on information technology in biomedicine, vol. 10, no. 4.
[17] R. Agrawal and J. Kiernan, "Watermarking relational databases," Proc. 28th International Conference on Very Large Databases (VLDB), Hong Kong, China, pp. 155–166, 2006.
[18] Francesc Sebe, Josep Domingo-Ferrer and Agusti Solanas, "Noise-Robust Watermarking for Numerical Datasets", V. Torra et al. (Eds.): MDAI 2005, LNAI 3558, pp. 134–143, Springer- Verlag Berlin Heidelberg 2005.
[19] Solachidis and I. Pitas, "Watermarking polygonal lines using Fourier descriptors," IEEE Comput. Graph. Appl., vol. 24, no. 3, pp. 44–51, May/Jun. 2004.
[20] C. G. Atkeson, A. W. Moore, and S. Schaal, "Locally weighted learning," Artif. Intell. Rev., vol. 11, pp. 11–73, Feb. 2006.
[21] Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, "Digital Watermarking and Steganography", Burlington, USA: Morgan Kaufmann Publisher, 2008.
[22] M. Potdar, S. Han, and E. Chang, "A survey of digital image watermarking techniques", in 3rd IEEE International Conference on Industrial Informatics (INDIN '05), pp. 709 - 716, Aug. 2005.
[23] G. Zeng and Z. Qiu, "Image watermarking based on dc component in DCT," in Proceedings of the 2008 International Symposium on Intelligent Information Technology Application Workshops, (Washington, DC, USA), pp. 573-576, IEEE Computer Society, 2008.
[24] C. Song, S. Sudirman, M. Merabti, and D. Llewellyn-Jones, "Analysis of digital image watermark attacks," in Proceedings of the 7th IEEE Consumer Communications and Networking Conference (CCNC 2010), pp. 1 - 5, Jan. 2010.
[25] S. P. Mohanty, R. Sheth, A. Pinto, and M. Chandy, "Cryptmark: A novel secure invisible watermarking technique for color images," in Proceedings of the IEEE International Symposium on Consumer Electronics (ISCE '07.), pp. 1 - 6, June 2007.
[26] K.Sridhar, Dr. Syed Abdul Sattar , Dr. M Chandra Mohan," Comparison of Digital Watermarking with Other Techniques of Data Hiding", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (1) , 2014, 350-353
[27] Prabhishek Singh, R S Chadha,"A Survey of Digital Watermarking Techniques, Applications and Attacks",International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 9, March 2013
[28] Bhupendra Ram," Digital Image Watermarking Technique Using Discrete Wavelet Transform And Discrete Cosine Transform", International Journal of Advancements in Research & Technology, Volume 2, Issue4, April-2013 ISSN 2278-7763
[29] Yong Zhu, Xiaohong Yu, Xiaohuan Liu, "An Image Authentication Technology Based on Digital Watermarking" International Conference on Sensor Network Security Technology and Privacy Communication System (SNS & PCS), 2013-IEEE.
[30] Lalit Kumar Saini, Vishal Shrivastava,"A Survey of Digital Watermarking Techniques and its Applications", International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 3, May-Jun 2014.

