ISSN:2249-5789
Ramji Shinde et al , International Journal of Computer Science & Communication Networks,Vol 4(6),214-215

Data Leakage Detection using Fake Objects for Suspected Users

Ramji Shinde[1], Prof. Amar Buchade[2]
Student P.I.C.T.pune.,P.I.C.T Pune
Ramshinde651@gmail.com,amar.buchade@gmail.com

Abstract

Nowadays data leakage is common across industries, academic and government offices. For the business purpose or research purpose the data must be shared among the different enterprises or agents. Once data is given to the agents, it should not reach to the unauthorized person. If somebody among the agents leaks the data, it may lead to great loss. To avoid this loss it is necessary to detect the leakage and stop doing business with that agent. Currently, techniques based on watermarking are used to detect the guilt of agent, but it has to modify the data which is not allowed in some cases.

The proposed work focuses on the novel method of detecting data leakage more efficiently based on Data allocation strategy using addition of fake objects in the original data

1. Introduction

Most of the users register him/her self with the trusted parties for research purpose or business purpose but due to data leakage problem they may suffer from great loss. For example we have a data distributer who has the data records of the entire customer also we have data agents which request the for research or business purpose. The agent may distribute data to other agents those who are not authorized which is not allowed. Hence our aim is to detect the guilty agent and find the guilt of that agent with respect to data leakage.

Most of the time we found the sensitive data in unauthorized or we may come across the unwanted things like phone call or advertisement on email. Due to this problem our data must be secure such that none of our trusted agents have cheated us. Our work is mainly focus on data distribution strategy which distributes the data among the agents and improves the probability of finding guilt of agent.

We have considered the situation where original data cannot be perturbed. Perturbation means we can modify the given data and make data less sensitive like replacing the values by range of number or add random noise to certain attributes.

We start in section II by literature survey. In section III we proposed an idea.

2. Literature survey

Nowadays data leakage is the common problem faced by all the data distributer. For the business purpose or research purpose the data must be shared among the different enterprises or agents. Once data is given to the agents, it should not reach to the unauthorized person. If somebody among the agents leaks the data, it may lead to a great loss. To avoid this loss, it is necessary to develop a method detecting the leakage and stop doing business with that agent. Currently, techniques based on watermarking are used to detect the guilt of agent which is also used in video and image watermarking, but it has to modify the original data which is not allowed in some cases where data is more sensitive.

Technique discuss in [4] is watermarking technique to find the guilt of agent. This is the traditional method in this method a unique code is inserted to the distributed copy. Once this copy is found in unauthorized place distributer will stop doing business with him for next time onward. Watermarking should not affect the quality of video or image. Mostly used for digital media like video or image. Algorithm should be capable to detect watermark after some common operations like signal processing.

Algorithm in [3] gives a good solution but it has limitations like domain specific. In this paper author discussed that Probability of finding guilt is similar to fault tolerant system. System failure is case where objects are guessed by the unauthorized agents and not leaked by any authorized agent.

Technique discuss in [6] is require the prior knowledge on the way a data view is created out of data sources.

Algorithm presented in [11] is applicable for the sample request or the explicit request but the paper has not discussed more about the shared data and different data allocation strategies. In this method fake object is added for every agent's data will increase the time complexity of the algorithm. In this paper it was assumed fixed set of agents with prior know request from agents.

Shabtai et al discuss about the method which is mostly useful for the shared objects. Also discuss algorithm which is heuristic in nature to calculate the guilt of an agent who have leaked the data from given data set. Running time of the proposed algorithm is very high and as it is in heuristic in nature it will not always produce the optimal result. For example adding fake email-id, and then if we receive emails from add company which we have not given email-id to them then we can find the

data leakage or the agent who leak the data. But in some cases fake object may also create problem like if we made changes in medical record then someone will be wrongly treated with wrong medicine. So in this case we can add one more fake record instead of modifying the exiting medical record.

Most of the technique discuss above depends on modification of original data. No one had discussed about the shared object distribution. In which the agent who are having same objects are equally treat as guilty agent.

3. Problem setup

Let us consider distributer have data set with n records (objects).

R= {R1, R2,…., Rn}.

Want to share objects with n number of agents

A= {A1, A2, A3,….., An}.

Request from agents are of two ways

1. Sample request = SAMPLE{R, m} any subset of mi records from R can be given to Ai
2. Explicit request = EXPLICIT{R, condition} on given condition Ai receives all objects that satisfy condition.

Constrains:

1. Leaking of one object is not related to other objects
2. If objects are guessed by unauthorized agent then we will not consider any leakage from agents.

For example, R = {r1, r2} and there are two agents with explicit data request such that

A1 = {r1, r2} and A2 = {r1} the value of sum objective in this case is

$$\sum_{i=1}^{z} \frac{1}{xi} \sum_{\substack{j=1 \\ j \neq i}}^{z} |xi \cap xj|$$

4. Proposed idea

We are proposing a novel method of data leakage detection such as with fake objects and different data agents we can find the guilt of agent. In this senior we will consider our algorithm such that once we distribute the objects to different agents; only to the suspected agents we will add fake objects to the data set. Distribution of data objects should be done such that probability of finding agent is guilty or not will be more.

5. Conclusion

On given data set probability of finding given agent is guilty or not is very important problem, depending on that data distributer will decide to do the business with that particular agent or not. For this agent should not identify that fake object is created by the distributer.

3. Reference

[1] Ruanaidh, JJK Ó., W. J. Dowling, and F. M. Boland. "Watermarking digital images for copyright protection." IEE Proceedings-Vision, Image and Signal Processing 143.4 (1996): 250-256.

[2] Hartung, Frank, and Bernd Girod. "Watermarking of uncompressed and compressed video." Signal processing 66.3 (1998): 283-301.

[3] Buneman, Peter, Sanjeev Khanna, and Tan Wang-Chiew. "Why and where: A characterization of data provenance." Database Theory—ICDT 2001. Springer Berlin Heidelberg, 2001. 316-330.

[4] Agrawal, Rakesh, and Jerry Kiernan. "Watermarking relational databases." Proceedings of the 28th international conference on Very Large Data Bases. VLDB Endowment, 2002.

[5] Sweeney, Latanya. "Achieving k-anonymity privacy protection using generalization and suppression." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 571-588.

[6] Cui, Yingwei, and Jennifer Widom. "Lineage tracing for general data warehouse transformations." The VLDB Journal—The International Journal on Very Large Data Bases 12.1 (2003): 41-58.

[7] Sion, Radu, Mikhail J. Atallah, and Sunil Prabhakar. "Rights protection for relational data." Knowledge and Data Engineering, IEEE Transactions on 16.12 (2004): 1509-1525.

[8] Guo, Fei, Jianmin Wang, and Deyi Li. "Fingerprinting relational databases." Proceedings of the 2006 ACM symposium on Applied computing. ACM, 2006.

[9] Czerwinski, Steve, Richard Fromm, and Todd Hodes. "Digital music distribution and audio watermarking." UCB IS 219 (2007).

[10] Buneman, Peter, and Wang-Chiew Tan. "Provenance in databases." Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM, 2007.

[11] Papadimitriou, Panagiotis, and Hector Garcia-Molina. "Data leakage detection." Knowledge and Data Engineering, IEEE Transactions on 23.1 (2011): 51-63.

[12] Hao, Fang, et al. "Protecting cloud data using dynamic inline fingerprint checks." INFOCOM, 2013 Proceedings IEEE. IEEE, 2013.

[13] Shabtai, Asaf, et al. "Optimizing Data Misuse Detection." ACM Transactions on Knowledge Discovery from Data (TKDD) 8.3 (2014): 16.