Novel User Level Data Leakage Detection Algorithm

T.Lakshmi Siva Rama Krishna, P.Bandavi, K.Priyanka, V.P.Vivek

Abstract— Data leakage detection (DLD) is the most widely used detection technique in many applications such as etc. detecting data leakage by various data sources is an important research issue. Several researchers contributed to detect the data leakage by proposing various techniques. In the existing DLD techniques the performance metrics such as accuracy and time have been neglected. In this paper, we have proposed a new DLD algorithm and named it as novel user level data leakage detection algorithm (NULDLDA). In the proposed NULDLDA we have considered the user point of view to know the leakage of data by which agent among several existing agents. We have implemented and compared the NULDLDA with existing DLD. The experimental results indicate that proposed NULDLDA improved the performance over DLD with respect to time and accuracy.

Key Terms: - IT; watermarking guilty agent; explicit data; DLP (data leakage prevention)

I. INTRODUCTION

Currently it is very important that should provide the prevention of data leakage within the system. In every organisation data leakage is the most dangerous issue. Data leakage infers unapproved transmission of unstable data or information from inside an organization to an outside objective where the protection of information is deal. A run of the mill procedure is to screen the data away and transmission for reveal sensitive information. In like manner it consider all data sensitive and perform distinguishing proof errand for every last one of those data. At any rate this makes the disclosure technique troublesome and distinguishing proof time to increase. Besides, the data proprietor may require giving disclosure reply to the DLD provider. However, there is likelihood that the provider cans detection the sensitive data. In order to constrain the leakage of the sensitive data, affiliation needs to keep clear text sensitive data from appearing in the limit. A screening instrument is use to check the records. As such one need another data acknowledgment course of action that empower provider to check the substance for gap without learning data. Thusly one need systems that gives exact disclosure with unassuming number of false alarm under various gap circumstance.

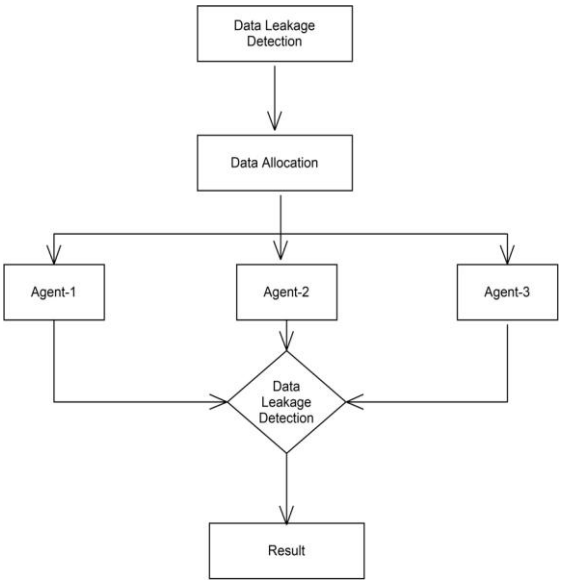

Figure: 1, Process of Data Leakage Detection

The proposed system mainly concentrated on the data leakage detection occurs at the various places and it is important to find the exact location of exact agent where the data leakage detection occurs. From the research point of view the fundamental occupation in purpose behind data incident among various data leak. There is diverse system to distinguish the data leak cause by human mistakes and keep the data by delivering an alert. Among various techniques, detect the data which is transmit for reveal of unstable data is ordinary. Furthermore it consider all data unstable and perform acknowledgment action for every last one of those data. However this makes the detection process difficult and detection time to increase.So there is a need of new data area game plan that empower provider to channel the substance for break without learning data. As such one need procedures that gives correct area with humble number of false caution under various break circumstance and result exhibits that the system upgrade the detection time. All together upgrade the detection time and ID of unstable data package, have helped framework is used which checks the repeat of occasion of data. Exceedingly isolated characteristics are considered as sensitive and fingerprints are made for them. Reiterated values are slighted in this methodology.

Quantifiable philosophy is use to make sensitive data and it is secured in table. The fingerprints are created by data spill area (DLD) provider and perceive potential breaks by organizing the fingerprints. The potential opening contains certifiable breaks and rackets with the objective that no one can get right information about the sensitive data. By then data proprietor post process data send by the DLD provider to check where there is any break in the sensitive data The objectives are to improve the recognizable proof time and to upgrade disclosure of unstable package.

II. MODULES

A. Data Allocation Module:

The vital purpose of our project is the data allocation issue as how can the distributor "intelligently" provides data to agents so as to update the chances of detecting a guilty agent, Admin can send the files to the authenticated client, clients can change their record subtleties, and so forth. Agent sees the secret key details through mail. So as to develop the chances of detecting agents that leak data.

B. Fake Object Module:

The distributor makes and adds fake articles to the data that he circulates to agents. Fake articles are objects produced by the distributor so as to expand the chances of detecting agents that leak data. The distributor might have the capacity to add fake items to the conveyed data so as to enhance his viability in detecting guilty agents. Our utilization of fake items is motivated by the utilization of "trace" records in mailing records. On the off chance that we give the wrong secret key to download the record, the duplicate file is opened, and that phony subtleties additionally send the mail. Ex: The fake item details will show.

C. Optimization Module:

The Optimization Module is the distributor's data allocation to agents has one requirement and one target. The agent's requirement is to full-fill distributor's requests, by providing them with the quantity of items they ask for or with every accessible article that full-fill their conditions. His goal is to have the capacity to distinguish a agent who releases any part of his data. Client can ready to bolt and open the documents for secure.

D. Data Distributor Module

A data distributor has given sensitive data to a lot of probably confided in agents (outsiders). A portion of the data is leaked and found in an unapproved put (e.g., on the web or someone's PC). The distributor must survey the probability that the leaked data originated from at least one operators, instead of having been autonomously assembled by different methods Admin can ready to see the which document is leaking and fake client's details also.

E. Agent Guilt Module

On the off chance that this individual can discover say 90 emails, we can sensibly figure that the likelihood of discovering one email is 0.9. Then again, if the articles being referred to are financial balance numbers, the individual may just find say 20, prompting an estimate of 0.2. We call this estimate pt, the likelihood that object t can be speculated by the objective. To rearrange the recipes that we present in whatever remains of the paper, we accept that all T objects have the equivalent pt,which we call p. Our conditions can be effectively summed up to differing pt's however they turned out to be bulky to show. Next, we make two presumptions with respect to the relationship among the different leakage occasions. The primary suspicion essentially expresses that an operator's choice to release an item isn't identified with different articles.

III. RELATED WORK

XiaokuiShu et al. in [1], has proposed fuzzy fingerprint technique. In this methodology DLD provider use interesting plan of the sensitive data digests. Set intersection point system is used between the procedure. In any case, set intersection point is an orderless as asked for of process isn't separate each time it may bungle. so now and again it deliver false caution rate. A.Broder et al. in [2], have proposed the Bloom Filter. Bloomfilter is data structure which is space capable and used for creating a set to help the support questions .Bloom filters allow false positives anyway space hold reserves much of the time have more load than shown. H. Yin, D. Tune et al. in [3], Propose a system Panorama as malware is extended of late. Malware is detected by getting real characteristic. In the examination, Panorama successfully detected all malware test anyway had a few false rate for exploring dark code test. G. Karjoth et al. in [4], privacy policy specification is hotbed of the investigation as use of web is extended in progressing years.The number of customer partaking in online development is extended. P3P and EPAL is use to address the privacy policies decided in quality criteria of programming need assurance. K. Edges et al. in [5],introduce Storages Capsule. It is used to protect private record on PC. Limit Capsule use cryptographic key to encoded report. with the objective that it will keep the grouping of the data. Checkpoint of the present system state are use to keep the track in like manner devastating device yield is use to achieve the goal already empowering access to limit compartment But it don't rely upon high uprightness. A. Nadkarni and W. Enck in [6], propose aquifer for preventing accidental information introduction in current working system structure and system. In aquifer, the entire UI work process is protectd using secret imprisonment that describe by the application developers . Nevertheless, it has nonattendance of usage seperation . Y. Jang et al. in [7], have proposed a way to deal with catch the structure lead that matches with increasingly indulgent semantics of the customers objective. In this procedure content based applications is used for discernment. In perspective of this idea, they have realized of model called s Gyrus2.

It will get the user intent .regardless, it won't get the period of event created by user. X. Shuet al.in [8], present a framework based data leak detection technique. Data proprietor makes not expect a basic showing with regards to in this methodology as it uses procedure to perceive the sensitive data. To evaluate the privacy for fingerprint framework they give a quantifiable system.

IV. DATA LEAKAGE DETECTION

Many researchers have been done research on the data leakage detection in various companies and it leads to prevention. Regardless, the significance of the data leaking or information leak expectation is the technique of substance monitoring and protecting them from the maltreatment [9]. Notwithstanding the way that examines on data leakage prevention are rising, there is little research on the disclosure of data leakage from the perspective of user behaviour [10]. Authors in [11] surveyed the DLP approaches and its issues with the best possible definition. The DLD and DLP process contains three phases, for instance, the data collection phase, analysis phase and the remedial action phase. The data gathering is beginning with the user internet or intranet logs and the database sources. The assembled data's are transported in the DLD and DLP examination arrange, which performs rule organizing, approach check, substance and setting affirmation frames. The context verification extracts the sender, source id, timings of the data access, association and size from the header information, etc., the substance is the pre-arranged data from typical verbalization and marking process. The detection and classifying the data into the predefined class reliably utilizes the security game plans and planning tests. Finally the DLD and DLP designs settle the issue by assurance appropriate recuperating exercises like alerted, blocking, allowing and doing some unique exercises in the security approach rule set. The detection and prevention attributes are analysis, detection and remedial actions. The data set technique consolidates the web online data's or in movement data's, user get to rights and separated recorded data from the database. These three sorts of data's are regularly assembled for the DLP technique. The data types used for the DLP. Data in movement is the data being transmitted edge one node to another inside a comparative framework or particular frameworks. Data being utilized is the data, which is available to the users in file design or email sorts out inside the applications. The data being utilized arrangement isn't mixed and this can be successfully deciphered. The third database data's are normally composed and anchored with strong access controls. The content based DLP screens sensitive data using standard verbalizations by perceiving the structure. For example account number, phone number and other sensitive nuances can be checked. As shown by this sort, authors in [12] proposed a DLP with typical explanation. Regardless, the framework was not productive and makes high false positive rates. The DLDs and DLPs are performed by master who can change the accessibility for the mystery data. Authors in [13] perceived different leaking channels, in which the data can be trade. This consolidates the helpful Medias like USB, memory cards and many. Authors surveys the activities related to the sentive data get as far as possible as demonstrated by the audit report. In figure-3, the data leakage detection is occurred at the user level is shown. Which has differing kinds of systems to turn away data leakage and keep up the ownership and detection using behavioural examination, etc., the most notable approaches under DLP is the use of cryptographic and watermarking techniques; this evades the data from the unapproved users. In the detection strategy, the data and behavioural examination with substance mining has a couple of enhancements.

Ensemble Algorithm

Step 1: Calculate total fake records as sum of fake records allowed.
Step 2: While total fake objects > 0
Step 3: Select agent that will yield the greatest improvement in the sum objective
Step 4: Create fake record
Step 5: Add this fake record to the agent and also to fake record set.
Step 6: Decrement fake record from total fake record set.
Algorithm makes a ensemble choice by selecting the agent that will yield the greatest improvement in the sum objective.

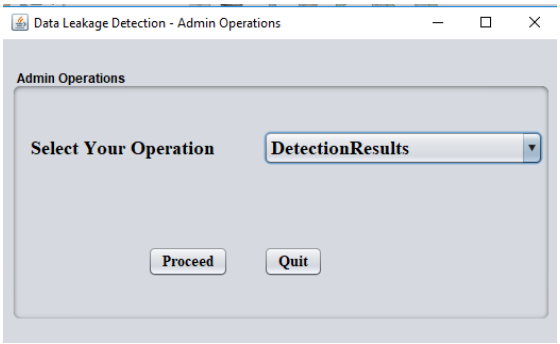

Figure: 2, Detection of Data Leakage

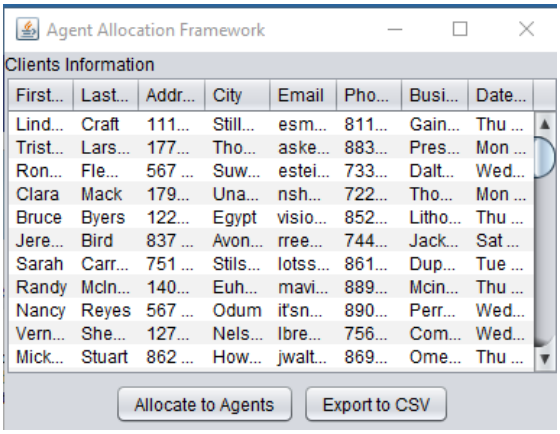

Figure: 3, finding the data leakage at the user level.

	ES	PS
Accuracy	70%	97%
Time(Sec)	10.98	03.1

Table: 1 Data leakage Detection based on the detection accuracy and time.

V. CONCLUSION

Data leakage is a burning issue in the field of data security and for all the companies who wants to secure the data. There are different investigates from various domains are continuously moving toward making data leakage detection and prevention procedures to direct this issue. Securing secrete and sensitive information is progressively basic. In this paper, the proposed ensemble data leakage detection identified the data leakage occurred at the agent level and shows the performance of the proposed system.

REFERENCES

1. DanfengYao ,XiaokuiShu and Elisa Bertino, Fellow IEEE,,"Privacy-Preserving Detection of Sensitive Data Exposure", IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 10, NO. 5, MAY 2015
2. A. Broder and M. Mitzenmacher, Network applications of bloom filters: A survey,Internet Math., vol. 1, no. 4, pp. 485509, 2004..
3. H. Yin, M. Egele, D. Song, C. Kruegel, and E. Kirda, "Panorama: Capturing system wide information flow for malware detection and analysis", in Proc. 14th Association for Computer Machinery Conf. Comput. Commun.Secur., 2007, pp. 116127.
4. G. Karjoth and M. Schunter, 'A privacy policy model for enterprises', in Proc. *15th IEEE Comput.Secur. Found. Workshop*, Jun. 2002, pp. 271281.
5. K. Borders, B. Lau, , E. V. Weele and A. Prakash, "Protecting confidential data on personal computers with storage capsules", in Proc. 18th USENIX security
6. Symp., 2009, pp. 367382.
7. A. Nadkarni and W. Enck, Preventing accidental data disclosure in modern operating systems, in Proc. 20th Association for Computer Machinery Conf. Comput.Commun.Secur., 2013, pp 10291042.
8. Y. Jang, B. D. Payne,, S. P. Chung and W. Lee," Gyrus: A framework for user-intent monitoring of text-based networked applications", in Proc. 23rd USENIX
9. Security Symp., 2014, pp. 7993.
10. X. Shu and D. Yao, *Data leak detection as a service*, in Proc. 8th Int. Conf. Secur. Privacy Commun.Netw., 2012, pp. 222240.
11. MogullR.Understandingandselectingadatalossprevention solution.Retrieved from ⟨https://securosis.com/assets/library/reports/DLP-Whitepaper.pdf⟩; 2010.
12. Boehmer, Wolfgang. "Analyzing Human Behavior Using Case-Based Reasoning with the Help of Forensic Questions." In *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*, pp. 1189-1194. IEEE, 2010.
13. Shabtai, Asaf, Yuval Elovici, and LiorRokach."*A survey of data leakage detection and prevention solutions"*.Springer Science & Business Media, 2012.
14. Yu, Fang, Zhifeng Chen, YanleiDiao, T. V. Lakshman, and Randy H. Katz. "Fast and memory-efficient regular expression matching for deep packet inspection."In *Architecture for Networking and Communications systems, 2006.ANCS 2006. ACM/IEEE Symposium on*, pp. 93-102. IEEE, 2006.
15. Hackl, Andreas, and Barbara Hauer."State of the art in network-related extrusion prevention systems."*Proceedings, 7th international symposuim on database engineering and applications* (2009): 329-35.
16. Y. Shapira, B. Shapira, and A. Shabtai. Content-based data leakage detection using extended fingerprinting,2013.
17. T. F. Gharib, M. M. Fouad, A. Mashat, and I. Bidawi,Self organizing map-based document clustering using WordNet ontologies Int. J. Com- put. Sci. Issues, vol. 9, no. 1, pp. 8895, 2012.