

Data Leakage Detection and Prevention of Confidential Data

Shubhangi G. Dhawase^{1*}, Bhagyashri J. Chaudhari², Neha S. Kolambe³, Poonam S. Masare⁴

^{1,2,3,4}Computer Science, SSBT's College of Engineering and Technology, Jalgoan, India

*Corresponding Author: sgdhawase@gmail.com

Available online at: www.ijcseonline.org

Accepted: 07/Jun/2018, Published: 30/Jun/2018

Abstract — Data leakage is big security threat to organization, when third party agent carried out data leakage. Perturbation method use for the detecting and preventing data leakage and also for assessing data third party agent. Firstly, watermarking techniques used for data leakage but in that techniques modification to original data take place. To overcome this disadvantage data allocation strategies is used. It improve feasibility of finding guilty agent. Data based on sample request and explicit request using allocation strategies allocate by the distributor to detect the agent. Fake object are given with original data. If one or more fake object is leaked, then distributor detect the leakage by the agent was guilty. So, perturbation is efficient technique handle data leakage and also make less sensitive to data before handle to agent.

Keywords— Agent, Distributor, Perturbation, Detection and Fake Object.

I. Introduction

When transmission of data is unauthorized from within organization to an external destination called as data leakage. In doing business, data must be handed to trusted agent. Distributor has given sensitive data to third party agent. Due to that some sensitive data is leaked during sending the data and found at unsafe place where, it has no any authorization. Distributor must access that data come from one or more agent and opposed to data leakage [22].

Section I contains the introduction of data leakage, background and motivation of data leakage detection. Section II contain the related work of data leakage detection, Section III contain the some measures of detection and security technique, Section IV contain the architecture of system, section V explain application, Section VI describes results and discussion of data leakage system, Section VII contain concludes research work with future directions. Reference explain in section VIII.

A. Background

Previously, data leakage is detected by watermarking, e.g., It add unique code in the document which is distributed. If that document is found at unauthorized party, leaker can be identified. Watermarking can be efficient but in that modification in the original data can be found. Watermark can also be destroyed, if recipient of data is malicious e.g. Company have partnership with other company so require data sharing. Data must be handle by many company and also handle by third parties agent [21].

B. Sensitive Data

Sensitive data is important information and it needs to be handle carefully. The reveal of sensitive data in storage and transmission creates a serious threat to organizational and personal security. Human misstep is the primary reason for data leak. The advantage of this system is without disclosing the sensitive data, information owner safely assign the detection operation to a DLD(data leak detection) provider i.e. distributor [10] [22].

C. Motivation

The motivation for proposed solution lies in problem with data leakage. Sometimes, it is needs to be detect and prevent the loss data from leaked and being stolen from the organization to the outside world. For that purpose many leakage technique are proposed but it is essential to detect the leakage of sensitive data as soon as possible before leaving the trusted network. So these basic problems in the data leakage must be overcome to reduce complexity and develop data leakage detection and prevention using perturbation technique [15][21].

II. Related Work

XiaokuiShu et al. in [1] presented fuzzy fingerprint technique that enhances data privacy during data leak detection operations. The data owner preprocess and prepares fuzzy fingerprints and release the fingerprints to DLD provider. The DLD provider computes fingerprints from the network traffic and identifies potential leaks in them.

Jing Zhang et al. in [2] described sequence alignment techniques used for detecting complex data leak patterns. Exposure of sensitive data is challenging task due to data transformation. Transformations (such as insertion and deletion) result in highly unpredictable leak patterns.

Michael Backes et al. in [3] illustrated watermark technique which enforces accountability by design. This helps to overcome the existing situation where most lineage mechanisms are applied only after a leakage has happened. They present an accountable data transfer protocol to transfer data between two entities. To deal with an entrusted sender and an entrusted receiver scenario associated with data transfer between two consumers, the protocols employ an interesting combination of the robust watermarking, oblivious transfer, and signature primitives. Cox algorithm is used for watermarking.

P. Papadimitriou et al. in [4] presented unobtrusive techniques for detecting leakage of a set of objects or records. They developed a model for finding the guilty of agents. They also present algorithms for sharing objects to agents, in a way that enhances the chances of identifying a leaker. Finally, choice of adding fake object also consider in distributed set. Such objects do not match to real entities but come into sight realistic to the agents. In a sense, here the fake objects act as a type of watermark for the entire set, without modifying any separate members. If one or more fake object given to agent is leaked then data leak detect and agent is guilty.

Subhashini Peneti et al. in [5] focus to data leakage prevention system with a time-stamp. In Data Leakage Prevention, the time stamp is very important for giving permission to access a particular data, as in a particular period of time the data is confidential after the time stamp the same data could be non-confidential. In time stamped based DLP two phases are there, Learning Phase and Detection Phase. In learning phase collect confidential and non-confidential documents of an organization. Then create clusters using K-means with cosine similarity function. For each cluster identify the key terms based on their frequency. Document is compared with the confidential score and time stamp in detecting phase, if the time stamp of the tested document is greater than or equal to the time stamp then that document is treated as a confidential and it is blocked.

Gilad Katz et al. in [6] proposed a new context-based model for accidental and intentional data leakage prevention is proposed. The context-based approach they proposed leverages the advantages of preventing data leakage by either looking for specific keywords and phrases or by using various statistical methods. Their new model consists of two phases: training and detection. During the training phase, they created clusters of documents. Then a graph representation of the confidential content of each cluster is generated. This representation consists of key terms and the context in which they need to appear in order to be considered confidential. During the detection phase, document tested is assigned to

several clusters. Its contents are then matched to each cluster's respective graph in an attempt to determine the confidentiality of the document.

Veroniki Stamati Koromina et al. in [7] aims to prevent the data leakage stemming from corporate email. When, an employee sends an email, which contains an attachment, from his corporate account to a recipient, the generated email is forwarded to the SMTP port which accepts outbound emails, on his system. SMTP proxy server can pick up the email and trigger the steganography scanner. Attachments are scanned and if they are clean the email is sent to main corporate server and finally send to intended recipient. If the attachment is not clean, i.e. a steganography payload is detected, alert for data leak can be triggered and that email will not be sent.

Yin Fan et al. in [8] researched a trustworthiness-based distribution model that aims at data leakage prevention. They study the application where there is a distributor, as a trusted party, managing and distributing files that contain sensitive information to authorized users when they require. In their model, first, based on the historical behaviors distributor calculate the distributors the user's trustworthiness based on his historical behaviors. Then according to the user's trustworthiness and his obtained file set overlapping leaked file set, the distributor accesses the probability of the user's intentional leak behavior as the subjective risk assessment. Then the distributor evaluates the user's platform vulnerability as an objective element. Finally, the distributor makes decisions whether to distribute the file based on the integrated risk assessment.

K. Borders et al. in [9] used Storage Capsule encrypted file container to Protect personal or confidential information's from personal computers and sent to the trusted module when the user finishes all their modification it the system is restored to its normal state and the output resumes normally.

Jason Croft et al. in [10] invented black-box differencing run two logical copies of the network, one with private data scrubbed, and compare outputs of the two to determine if and when private data is being leaked. To ensure outputs of the two copies match, build upon recent advances that enable computing systems to execute deterministically at scale and with low overheads. Building general-purpose schemes that leverage black-box differencing to mitigate leakage of private data could be useful approach using building block. This modified Linux kernel would not be an intrusive solution for network with high demands for security, a more system transparent design would be ideal necessary.

I. DETECTION AND SECURITY TECHNIQUE

A. Perturbation Technique

When the distributors sensitive data has been leaked by agents and to identify the agent that leaked the data. Perturbation method use for the detecting and preventing data leakage and also for assessing data third party agent which

makes the data modified and making it less sensitive before being handed to agents. This technique is proposed to develop for finding out the guilt agents. The algorithm uses for allocation of data request and objects to agents, in that way chances are improved for identifying a leaker. Distributor also has authority to add fake objects in the distributed set. These type of objects are not real entities but pretend as the real to the agents [13].

B. Key Generation

Secrete key is generated by the distributor and use by agent . Key is generated by the randomize algorithm with send file key is given and generated by randomly any number [18][20].

IV. ARCHITECTURE

In the architecture of data leakage detection and prevention includes the various blocks i.e. agent, database, distributor, fake agents and probability distribution of fake agents.

A. Block Diagram

In the block diagram there is the transfer of data and database stored the records of the fake agent for guilt model. It must calculate the probability through fake objects calculation. Sometimes there is the possibility of illegal transfer of data to fake agents [17].

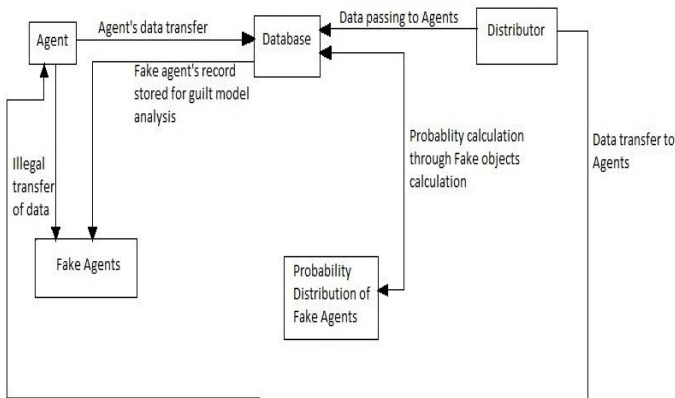


Figure 1. Architecture of Data Leakage Detection & Prevention

- 1) *Agent*: Agent transfer the important data to the database and it stored the record of data. Also there is the chances of illegal transfer of data to fake agents.
- 2) *Database*: Fake agent's record is stored in database for guilt model analysis.
- 3) *Distributor*: Distributor pass and transfer the data to agents.

- 4) *Fake Agents*: It is illegal agent i.e. fake agent that has no any authority to take data through database directly.
- 5) *Probability Distribution of Fake Agents*: In the database, probability calculation is through fake objects calculation.

B. Algorithm

```

Input:  $R_1, \dots, R_n, cond_1, \dots, cond_n, b_1, \dots, b_n, B$ 
Output:  $R_1, \dots, R_n, F_1, \dots, F_n$ 
1:  $R \leftarrow \emptyset$ 
2: for  $i = 1, \dots, n$  do
3:   if  $b_i > 0$  then
4:      $R \leftarrow R \cup \{i\}$ 
5:    $F_i \leftarrow \emptyset$ 
6: while  $B > 0$  do
7:    $i \leftarrow \text{SELECTAGENT}(R, R_1, \dots, R_n)$ 
8:    $f \leftarrow \text{CREATEFAKEOBJECT}(R_i, F_i, cond_i)$ 
9:    $R_i \leftarrow R_i \cup \{f\}$ 
10:   $F_i \leftarrow F_i \cup \{f\}$ 
11:   $b_i \leftarrow b_i - 1$ 
12:  if  $b_i = 0$  then
13:     $R \leftarrow R \setminus \{R_i\}$ 
14:   $B \leftarrow B - 1$ 
    
```

Where,

- $R_1, R_2, \dots, R_n \rightarrow$ User request for data
- $b_1, b_2, \dots, b_n \rightarrow$ Number of agents
- $F_1, F_2, \dots, F_n \rightarrow$ Number of fake Objects

Step 1 : The creation of a fake object for agent U_i as a black-box function $\text{CREATEFAKEOBJECT}(R_i; F_i; cond_i)$ that takes as input the set of all objects R_i , the subset of fake objects F_i that U_i has received so far and $cond_i$, and returns a new fake object.

Step 2: This function needs $cond_i$ to produce a valid object that satisfies U_i 's condition.

Step 3: Set R_i is needed as input so that the created fake object is not only valid but also identical from other real objects.

For example, Employee rank and a salary attribute may take into account the distribution of employee ranks, the distribution of salaries as well as the correlation between the two attributes that create function of fake payroll record.

Step 4: By introduction of fake objects key statistics do not change if agent using that type of statistics in their work.

Step 5: Finally, to ensure proper statistics function $\text{CREATEFAKEOBJECT}()$ aware of fake objects F_i added.

Step 6: The distributor can also use function CREATEFAKEOBJECT() when it wants to send the same fake object to a set of agents.

Step 7: In this case, the function arguments are the union of the Ri and Fi tables respectively, and the intersection of the conditions condi's.

Step 8: Although we do not deal with the implementation of CREATEFAKEOBJECT() we note that there are two main design options.

Step 9: The function can either produce a fake object on demand every time it is called, or it can return an appropriate object from a pool of objects created in advance.

As a conclusion it is made clear that fake objects can be anything depending on the distributor and these paper does not deal with creation of fake objects, but CREATEFAKEOBJECT() method is defined in order to distribute the objects to the agents with fake object or without fake object depending on the request.

C. Comparison Between Watermarking and Perturbation

Table 1. Comparison Table

NO	Parameter	Watermarking	Perturbation
1	Security	[a] It use the watermark that can destroy easily	[a] It use the Key function
2	Data Loss	[b] Data loss is high	[b] Data loss is low
3	Data Prevention	[c] No data prevention	[c] Prevention by block agent
4	Agent Identity	[d] It does not identify agent	[d] It identify the agent

In the table, it gives comparison of different parameters between the watermarking and perturbation. The first important parameter is security, in watermarking use the watermark that can easily destroy where in perturbation uses the key function. Data loss is high in watermarking whereas in perturbation is low. Data is prevented by block the agent in perturbation and watermarking has no any prevention. It does not identify the agent in watermarking but perturbation technique is identify the agent easily[23][27].

V. APPLICATIONS

- A company may have partnerships with other companies that require sharing customer data.
- 2) It helps in detecting whether the distributor's sensitive data has been leaked by the trustworthy or authorized agents.

- 3) A hospital may give records of patients to researcher who will suggest new treatments.
- 4) It helps to identify the agents who leaked the data.
- 5) Reduces cybercrime.
- 6) Used for transfer the student's record between the faculty members in the school or college.
- 7) A enterprise may outsource its data processing, so data must be given to various other companies.

VI. RESULT & ANALYSIS

There are two login windows, one for distributor and one for agent, each have name and password. First distributor is login, after login of distributor. Distributor have five pages which is home, send file, view file, leak file and logout. After that agent is login on the window.



Figure 2. Distributor Window

- In figure 2, after login to distributor. Distributor have five pages which is home, send file, view file, leak file and logout.

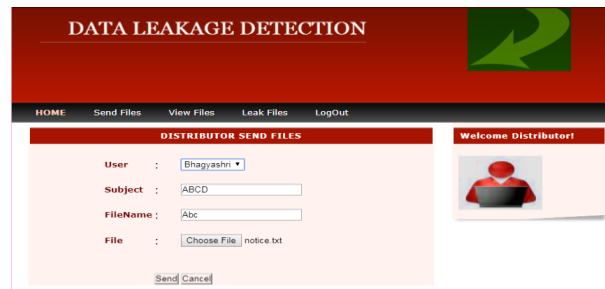


Figure 3. Distributor Sending File to Agent

- In figure 3, if distributor wants to send a file then distributor goto send file page then give the name of agent, subject, filename, name of the file.



Figure 4. File Sending Successfully Window

- In figure 4, after sending file to agent it give login is successfully message and give the login to agent.

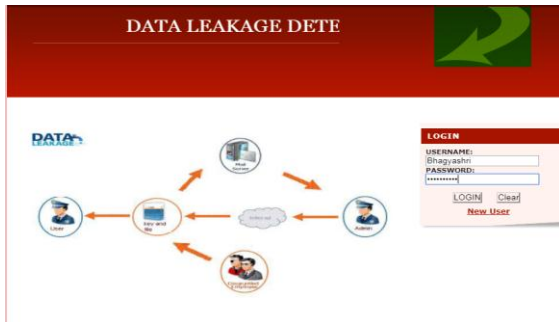


Figure 5. Agent Login

- In figure 5, login page of agent which has required name and password.

- In figure 8, after clicking on detail of file it give that window.

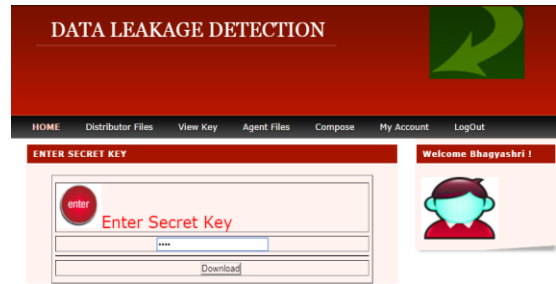


Figure 9. Enter Secret Key

- In figure 9, distributor file downloaded by key, to give the key window is appear.



Figure 6. Agent Receive Distributor File

- In figure 6, after login to agent then it give agent window. In that distributor file it give file name with details.

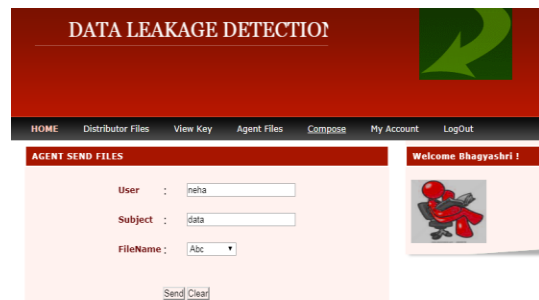


Figure 10. Compose File to Fake Agent

- In figure 10, if agent compose file to new agent it need new agent and send file it.

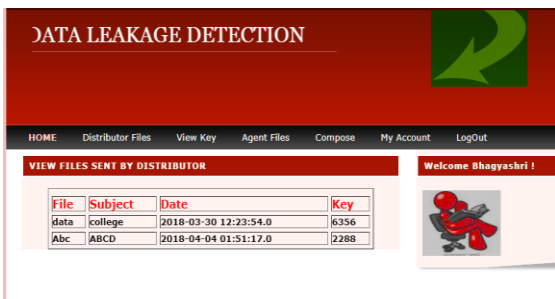


Figure 7: Details of File(key)

- In figure 7, to download distributor file agent need key. Key is given with file present in view key.

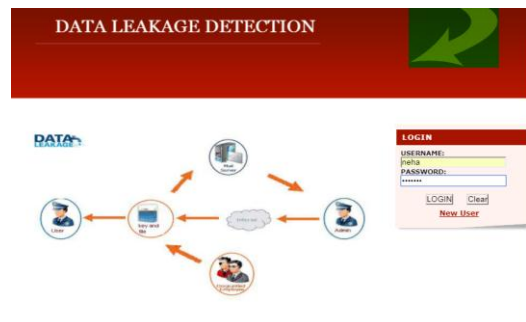


Figure 11. Login of Fake Agent

- In figure 11, login page to the new fake agent.

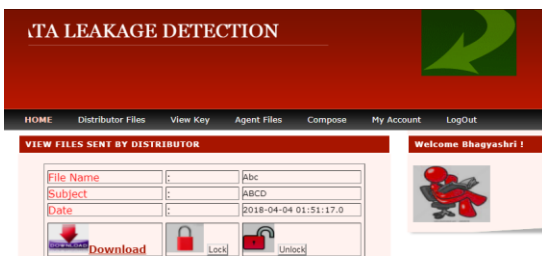


Figure 8. Download File

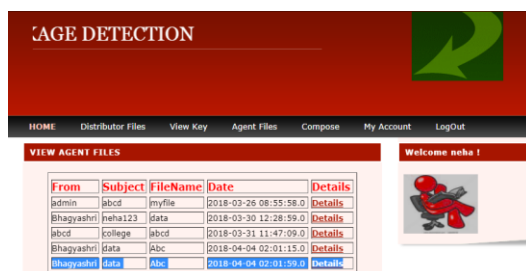
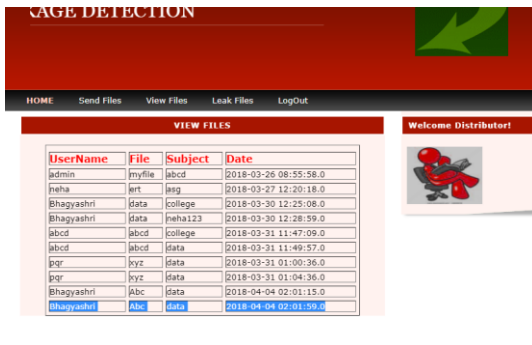


Figure 12. Receive Agent File

- In figure 12, the agent file new fake agent can see file.



UserName	File	Subject	Date
admin	myfile	abcd	2018-03-26 08:55:58.0
neha	ert	leg	2018-03-27 12:20:18.0
Bhagyaashri	data	college	2018-03-30 12:25:08.0
Bhagyaashri	data	neha123	2018-03-30 12:28:59.0
abcd	abcd	college	2018-03-31 11:47:09.0
abcd	data	data	2018-03-31 11:49:57.0
pr	pr	data	2018-03-31 01:50:36.0
pr	pr	data	2018-03-31 01:54:36.0
Bhagyaashri	abc	data	2018-04-04 02:01:15.0
Bhagyaashri	abc	data	2018-04-04 02:01:59.0

Figure 13. Detect Data Leak

- In figure 13, as the file is leak distributor login and check the leak file then it give the leak file.

VII. CONCLUSION

Rapidly increasing data leak cases in now a day taken into account and so to detect the leakages here proposed a perturbation technique for data leakage detection in organization's network traffic. The data leakage detection and prevention should ensure sensitive data remain safe due to data leakage detection and prevention. It identifying the probability of data leakage is dominant especially when the data is confidential and sensitive in nature. Firstly, leakage detection is handled by watermarking which modifies original objects before being transmitted for security reasons, system does not need any alteration of original objects. Perturbation is a very useful technique where the data is modified and made less sensitive and it also help for prevention.

REFERENCES

- [1] XiaokuiShu, Danfeng Yao and Elisa Bertino, "Privacy-Preserving Detection of Sensitive Data Exposure", IEEE Transactions on Information Forensics and Security, 1092-1103.
- [2] XiaokuiShu and Jing Zhang, Danfeng Daphne Yao and Wu Chun Feng, "Fast Detection of Transformed Data Leaks", IEEE Transactions on Information Forensics and Security, 528-542.
- [3] Michael Backes, Niklas Grimm and Aniket Kate, "Data Lineage In Malicious Environments", IEEE Transactions on Dependable and Secure Computing, 178-191.
- [4] P. Papadimitriou, H. Garcia-Molina, "Data Leakage Detection", IEEE Transactions On Knowledge And Data Engineering, 51-63.
- [5] SubhashiniPeneti and B. Padmaja Rani, "Data Leakage Prevention System With Timestamps", International Conference on Information Communication and Embedded Systems, 1-6.
- [6] Gilad Katz, Yuval Elovici, and BrachaShapira, "CoBAN: A context based model for data leakage prevention", Information science on Springer.

- [7] VeronikiStamatiKoromina and Christos Ilioudis, "Insider Threats in Corporate Environments: A Case Study for DLP", in Proc. ACM.
- [8] Yin Fan, Wang Yu, Wang Lina, YuRongwei, "A Trustworthiness-Based Distribution Model for Data Leakage Prevention", Wuhan university journal of natural sciences,2010, Vol.15 No.3, 205-209.
- [9] K. Borders, E. V. Weele, B. Lau and A. Prakash, "Protecting Confidential Data on Personal Computers with Storage Capsules", 18th USENIX Security Symposium, 2009.
- [10] Jason Croft, Matthew Caesar, Xin Liu, and Wenjuan Gong, "Towards practical avoidance of information leakage in enterprise networks", International Journal of Distributed Networks,10 November 2011.
- [11] P.Buneman, S.Khanna, and W.C.Tan, "Why and Where: A Charaterization of Data provenance", Proc.Eighth Int'l Conf. Database Theory(ICDT '01'),J.V. den Bussche and V.Vianu,eds.,pp.316-330,Jan.2001
- [12] P.Buneman and W.C.Tan, "Provenence in Databases", Proc ACM SIGMOD, pp.1171-1173,2007
- [13] Y.Cui and J.Widom, "Lineage Tracing For General Data Warehouse Transformations", The VLDB J.vol.12,pp.41-58,2003.
- [14] J.J.K.O.Ruanaidh, W.J.Dowling, and F.M.Boland, "Watermarking Digital Images For Copyright Protection", IEE Proc. Vision, Signal and Image Processing,vol.143,no.4,pp.250-256,1996.
- [15] F.Hartung and B.Girod, "Watermarking of Uncompressed and Compressed Video", Signal Processing, vol.66, no.3,pp.283-301,1998.
- [16] S.Czerwinski, R.Fromm,and T.Hodes, "Digital Music Distribution and Audio watermarking", <http://www.Scientificcommons.org/43025658>,2007.
- [17] S.Jajodia, P.Samarati, M.L.Sapino,and V.S. Subrahmanian, "Flexible Support For Multiple Access ControlPolicies", ACM Trans. Database Systems vol.26.no.2,pp.214-260,2001.
- [18] P.Bonatti, S.D.C.di Vimercati,and P.Samarati, "An Algebra For Composing Access Control Policies", ACM Trans. Information and system Security,vol.5,no.1,pp.1-35, 2002.
- [19] L. Sweeney, "Achieving K-Anonymity Privacy Protection Using Generalization and Suppression", <http://en.Scientificcommons.org/43196131>, 2002.
- [20] R. Sion, M. Atallah, and S. Prabhakar, "Rights Protection for Relational Data", Proc. ACM SIGMOD, pp. 98-109, 2003.
- [21] Ensaf Hussein, Mohamed A. Belal, "Digital Watermarking Techniques, Applications and Attacks Applied to Digital Media: A Survey", IJERT, ISSN: 2278-0118, Vol. 1 Issue 7, September2012.
- [22] Sandip A. Kale, Prof. S.V.Kulkarni, "Data Leakage Detection", International Journal of Advanced Research in Computer and Communication Engineering, ISSN: 2278-1021, Vol. 1, Issue 9, November 2012.
- [23] Upasana Yadav, J.P.Sharma, Dinesh Sharma, Purnima K. Sharma, "Different Watermarking Techniques & its Applications: A Review", IJSER, ISSN 2229-5518, Volume 5, Issue 4, April-2014.
- [24] Cox, I.J.; Miller, M.L.; Bloom, J.A., "Digital Watermarking", Morgan Kaufmann, 2001.
- [25] Prabhishkek Singh, R S Chadha, "A Survey of Digital Watermarking Techniques, Applications and Attacks", IJEIT, ISSN: 2277-3754, Vol. 2 Issue 9, March-2013.