IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 3, March 2017

Privacy Preserving Sensitive Data Coverage: Design and Analysis

Deepika Patel[1], Dr. Pramod S. Nair[2], Mr. Rudresh Shah[3]

ME Student, CS Dept, Medicaps Institute of Technology & Management, Indore, India[1]

HOD, CS Dept, Medicaps Institute of Technology & Management, Indore, India[2]

Assistant Professor, CS Dept, Medicaps Institute of Technology & Management, Indore, India[3]

Abstract: Privacy and security has become major issues in today's communication. In respect to the last ten years the nature and utilization of the communication technology is changed much frequently. Due to this a significant amount of data is communicated in a fraction of seconds. Therefore the traditional computational techniques have moved towards the big data processing and analytics. In this environment the entire client module and administration is directly connected with the same data sources. Due to this the communication becomes easy but security and privacy concern in communicated data has appeared. Sometimes these issues are arising due to data leakage. In this presented work the main aim is to investigate the privacy and security concerns due to data leakage in big data environment. The main reason to utilize the big data is to demonstrate the real time system using twitter accounts to fetch and improve the sensitivity of data. During the investigation the promising approach is appeared where the sliding window and fuzzy logic based system is provided to analyze and reform the data. But this approach is found slow processing capability by which the system performance is affected. Due to this a new approach using the random walk technique is prepared by modification of existing system to enhance the resource consumption of the system.

Index: Privacy preserving, DLD, Sensitive Data, Security, Data Exposure.

I. INTRODUCTION

Privacy preserving is increasing in its importance since privacy becomes a major concern for both customers and enterprises in today's corporate marketing strategies. This raises challenging questions and problems regarding the use and protection of private messages, especially for context-aware web service. One principle of protecting private information is based on who is allowed to access private information and for what purpose [1] [2]. The exposure of sensitive data in storage and transmission poses a serious threat to organizational and personal security. Data leak detection aims at scanning for exposed sensitive data. Because of the large content and data volume, such a screening algorithm needs to be scalable for a timely detection. Development of online social networks and publication of social network data has led to the risk of leakage of confidential information of individuals. This requires the preservation of privacy before such network data is published by service providers

A. Categories of Privacy Breach

A privacy breach occurs when private and confidential information about the user is disclosed to an adversary. So, preserving privacy of individuals while publishing user's collected data is an important research area. The privacy breaches in social networks can be categorized into three types [3] [4]:

1) Identity Disclosure: Identity disclosure occurs when an individual behind a record is exposed. This type of breach leads to the revelation of information of a user and relationship he/she shares with other individuals in the network.

2) Sensitive Link Disclosure: Sensitive link disclosure occurs when the associations between two individuals are revealed. Social activities generate this type of information when social media services are utilized by users.

3) Sensitive Attribute Disclosure: Sensitive attribute disclosure takes place when an attacker obtains the information of a sensitive and confidential user attribute. Sensitive attributes may be linked with an entity and link relationship.

B. Why Data Needs Protecting?

With the advent of the Internet and new technologies that allow easier, quicker, as well as anonymous access to more information than ever before, people have now become more aware of identity theft and make conscious decisions on how to protect themselves. If the information is sensitive, it's likely to be protected by laws, regulations, or policies.

However, you can take an active approach to making sure your information has not fallen into the hands of those who would misuse it for financial gain or other reasons. Identity theft and online crime have now surpassed any other form of crime in profits earned, including drug-

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 3, March 2017

related crimes, so it is important to be wary of how exposed your information really is [5].

The remainder of this paper is organized as follows: Section II presents the related studies in which contribution of previous work have added. Section III proposes a purpose based access framework which includes detailed information of purposes and privacy preserved control evaluation. Section IV listed the work results with comparison of traditional approach the comparisons demonstrate the significance of the work in this paper. Finally, the conclusion of the paper and further work are given in Section V.

II. RELATED WORK

In [6] this paper introduced and demonstrated Revolver, a novel approach and tool for detecting malicious JavaScript code similarities on a massive scale. Revolver's approach is based on identifying scripts that are similar and catching into account an Oracle's classification of each one script. By doing this, Revolver can pinpoint scripts those have high sameness but are classified differently (detecting likely evasion attempts) and improve the accuracy of the Oracle.

This performed a massive scale evaluation of Revolver by running it in parallel with the popular Wepawet drive-by detection tool. This identified several cases of evasions that are used in the wild to evade this tool (and, likely, other tools based on similar techniques) and fixed them, improving this way the accuracy of the honey client. The advantage of this paper is this lack of application separation did not expose it as a concern. But here it may not be trusted with that data.

In [7], network-based data-leak detection (DLD) technique, have important feature of which is that the detection does not require the information proprietor to show the content of the sensitive data. Instead, only a small amount of specialize digests are needed. In comparison to host-based approaches, network-based data-leak detection focuses on examining the (unencrypted) content of outbound network packets information.

For example, a naive solution requires inspecting every packet for the occurrence of any of the sensitive data defined in the database. Such solutions generate alerts if the sensitive data is found in the outgoing traffic However, this naïve solution requires to store sensitive data in plaintext at the network interface, which is highly undesirable. This is not efficient enough for practical data leak inspection in this setting.

Batya Kenig and Tamir Tassa [8] introduced a method of mining closed frequent generalized records. Experiments show that the significance of algorithm is not limited to the theory of k-anonymization. This achieves lower information loss than the leading approximation algorithms. However the traditional k-anonymity models consider that all values of the attributes are sensitive and need to be protected. In fact, the values which will breach individual's privacy are in the minority of the whole sensitive attribute dataset. The previous models lead to excessive generalization and suppression that leads to more information loss in publishing data.

This paper Hua Wang et al. [9] proposes a purpose-based access control model in distributed computing environment for privacy preserving policies and mechanisms, and describes algorithms for policy conflicting problems. The mechanism enforces access policy to data containing personally identifiable information. The key component is purpose involved access control models for expressing highly complex privacy-related policies with various features.

A policy refers to an access right that a subject can have on an object, based on attribute predicates, obligation actions, and system conditions. Policy conflicting problems may arise when new access policies are generated that are possible to be conflicted to existing policies.
As a result of the policy conflicts, private information cannot be well protected. The structure of purpose involved access control policy is studied and efficient conflict checking algorithms are developed and implemented. Finally a discussion of our work in comparison with other related work such as EPAL is presented.

Traditionally, as soon as confidentiality becomes a concern, data are encrypted before outsourcing to a service provider. Any software-based cryptographic constructs then deployed, for server-side query processing on the encrypted data, inherently limit query expressiveness.

Here, Sumeet Bajaj et al. [10] introduce TrustedDB, an outsourced database prototype that allows clients to execute SQL queries with privacy and under regulatory compliance constraints by leveraging server-hosted, tamper-proof trusted hardware in critical query processing stages, thereby removing any limitations on the type of supported queries. Despite the cost overhead and performance limitations of trusted hardware, they show that the costs per query are orders of magnitude lower than any (existing or) potential future software-only mechanisms. TrustedDB is built and runs on actual hardware and its performance and costs are evaluated here.

III. PROPOSED WORK

The solution development methodology states the techniques graph theory for finding the optimum solution for sensitive data privacy preservation and text based search technique enhancement.

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 3, March 2017

A. Methodology

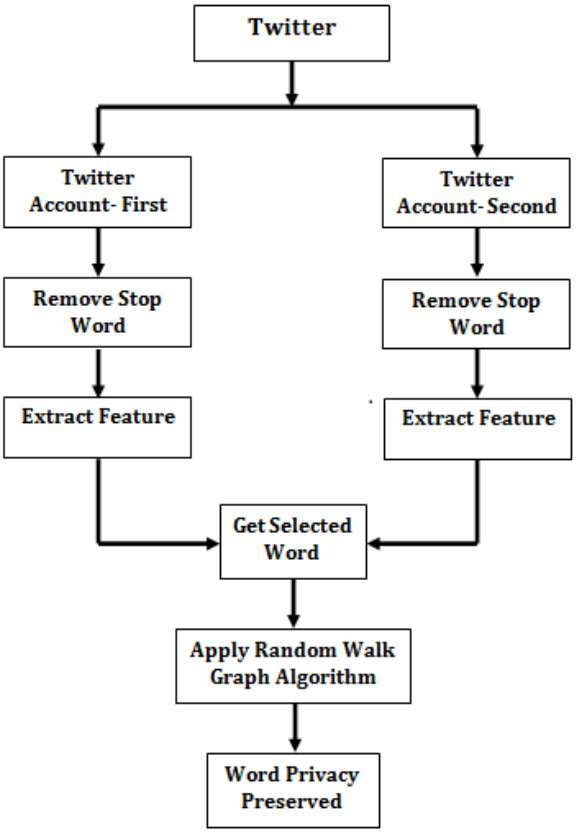

Fig 1: Flow Chart

In this work we are presenting a privacy preservation of the Sensitive private data that ensure detection of data it's been leak to publicly. In figure gives the details about the working scenario of the whole research work. Different phase with their description listed below

1) Twitter: Twitter is a social networking application which allows people to micro-blog about a broad range of topics. Micro-blogging is defined as "a form of blogging that lets you write brief text updates (usually less than 200 characters) about your life on the go and send them to friends and interested observers via text messaging, instant messaging (IM), email or the web". For the data privacy, we capture some tweets from the twitter account.

2) Twitter Account: In this phase we create account on twitter social networking site. For this we post some tweets on account and follow other peoples. By this, we used twitter data of the following peoples. In this proposed work make two twitter accounts i.e. first and second. To use this twits by means of account access.

3) Stop Word Removal: some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words. The general strategy for determining a stop list is to sort the terms by collection frequency and then to take the most frequent terms, often hand-filtered for their semantic content relative to the domain of the documents being indexed, as a stop list, the members of which are then discarded during indexing. Therefore, in this phase remove all this stop word for further processing.

4) Feature Extraction: In this phase, we extract the feature of sensitive data which we need to preserve when it is release. Sensitive data of peoples are need to protect while it been in general in social network or other places. Finally, to get the selected word we extract all essential features of this twitter data.

5) Random Walk Graph Algorithm: The stochastic process formed by successive summation of independent, identically distributed random variables – is one of the most basic and well-studied topics in probability theory. Given an undirected, connected graph G (V, E) with |V| = n, |E| = m a random "step" in G is a move from some node u to a randomly selected neighbor v. A random walk is a sequence of these random steps starting from some initial node.

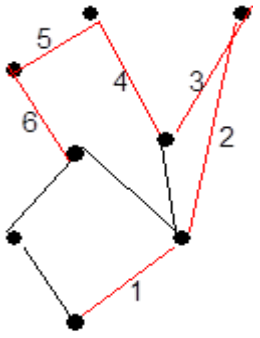

Fig 2: Random Walk

In this phase we apply Random walk Algorithm for further process. After taking selected row data that there we input this data to random walk algorithm. By applying this process we get privacy preservation of sensitive data.

B. Proposed Algorithm

This section introduces the summarized steps of the privacy preserving detection proposed for secured sensitive data information from public environment among social network.

TABLE I Privacy Preserving for Sensitive Data

Input: Twitter Data **Output: Preserved Data**
Process: **1:***AccessTweets from TwitterAccount* i. $twitter_{account} - first$ i. $twitter_{account} - second$ **2:***Remove StopWords from both TwitterAccount* **3:***ExtractFeatures from both TwitterAccount* **4:***Produced PreservedWord Information*

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 3, March 2017

5: *Apply RandomWalkAlgorithm*
i: *select preserved word*
ii: *find length of presereved word using MD5(W)*
iii: *generate any random number (R)*
iv: *compare word length and random number*
v: ***if*** *(W == R)*
asterisk all middle letterescap first and last
vi: ***elseif*** *(W! = R)*
repeat step 9
vii: ***endif***
6: *show presereved word differently*

IV. RESULT ANALYSIS

A. Time complexity

In general time complexity is the amount of time required to execute the algorithm for computing the outcomes. In this context the time consumption of the approach shows amount of time required to develop the release document after scanning of the documents. The time consumption of the proposed and traditional system is given using figure 3.

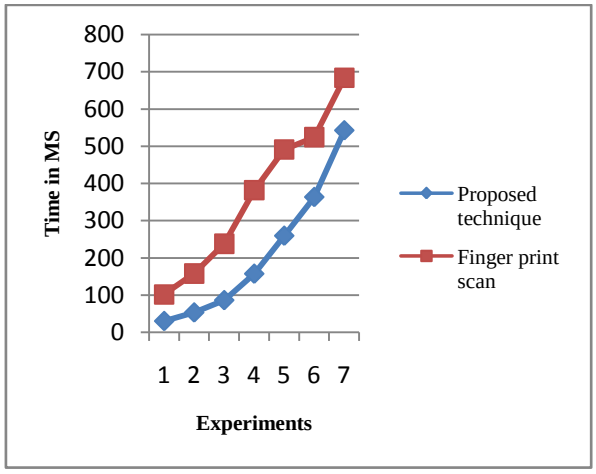

Fig 3: Time Consumption

The time consumption of both the techniques is given using figure 3. In this diagram the X axis shows the different experiments performed with the system and the Y axis shows the amount of time required for execution. To represent the performance of the algorithms the blue line shows the performance of proposed technique and the red color line shows the performance of the proposed technique. According to the obtained results the performance of the proposed system is much effective than the traditional approach. The key reason behind the less time consumption of proposed approach is their random walk process of search. Therefore the method guarantees about the search less than the liner search process. Additionally to justify the performance of both the algorithm the average time consumption of the system is also computed. To compute the mean time the following formula is used:

$$Mean\ time = \frac{1}{N}\sum_{i=1}^{N} time_i$$

The figure 4 shows the performance of both the algorithm in terms of mean time consumption. In this diagram the X axis contains the methods used for comparison and the Y axis shows the mean time required for developing similar documents. According to the obtained results the proposed technique consumes very fewer time as compared to the traditional technique.

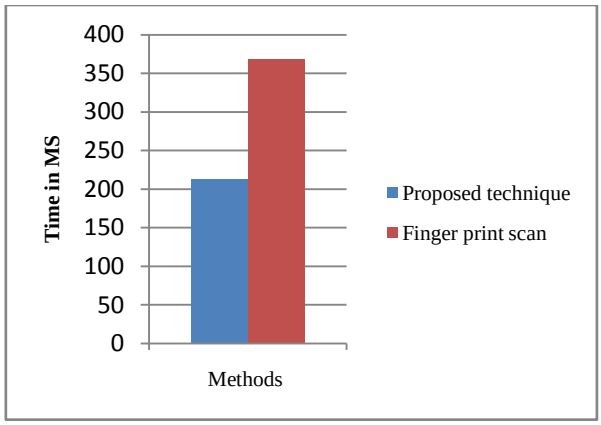

Fig 4: Average Time Consumption

B. Space complexity

The space complexity of any algorithm shows the amount of main memory required to execute the selected algorithm. The memory consumption of algorithm is also termed as the space complexity of the algorithm. The memory consumption of the proposed privacy preserving technique and the traditional finger print scan technique is demonstrated using figure 5. In the given diagram the X axis shows the different experiments executed with the different sets of twitter data and the Y axis shows the amount of memory consumed in terms of KB (kilobytes). According to the obtained experimental results both the techniques consumes similar patterns of memory. But the proposed technique requires less amount of main memory as compared to the traditional technique.

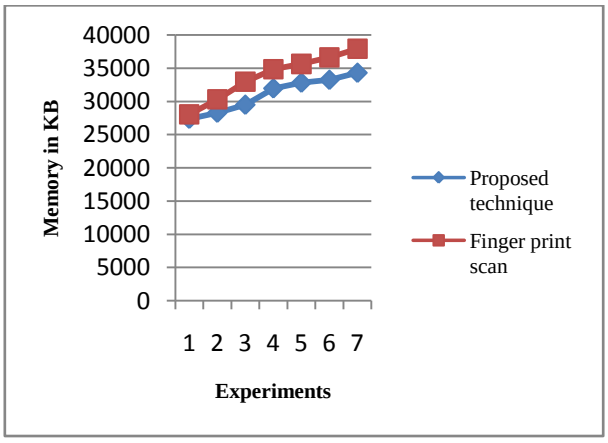

Fig 5: Memory Consumption

DOI 10.17148/IJIREEICE.2017.5334

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 3, March 2017

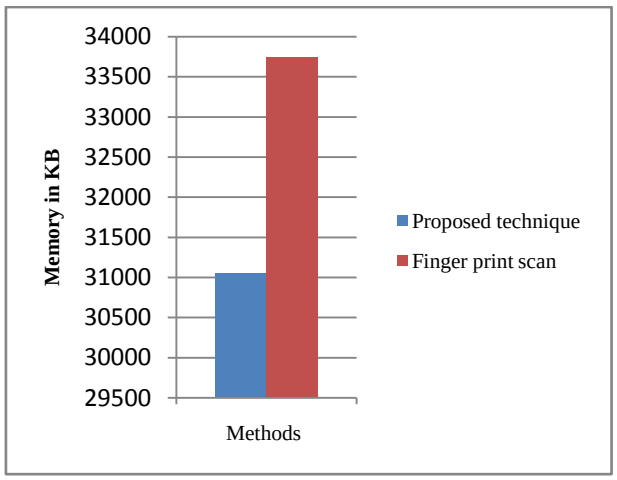

Fig 6: Mean Memory Consumption

To demonstrate the clear difference in memory consumption the mean memory consumption of the techniques is also computed using the following formula:

$$mean\ memory = \frac{1}{N}\sum_{i=1}^{N} memory_i$$

The figure 6 shows the mean memory consumption of the algorithms. According to the diagram the X axis contains the methods used and the Y axis shows the memory in KB (kilobytes). According to the results the proposed technique requires less memory as compared to traditional approach.

V. CONCLUSION

The privacy and sensitivity in digital data is a major concern in different applications. The applications now in these days not only work in a single PC (personal computer). These enabled for network usages also, therefore the data is communicated in public network form a secure environment. The public network is sometimes not much trustworthy for communication of private data. Thus for private communication with the end clients data privacy checks are implemented. These checks are used to cross verify the entire communicated text among two parties for preserving the private and sensitive contents in the public networks.

In order to prevent such kind of data leakage which is not intentionally communicated a privacy preserving filter is developed in this work. thus for demonstration the two twitter accounts are used, and before posting of the data on any one's wall the privacy filter is used. In previous works that filter is used for preserving the data for organizational aspects. In the previous models the sliding window technique is used to find out the target contents and then the alteration process is used to hide the obtained contains. These methods are accurate for finding contents but the main issue is their performance in terms of time complexity. Thus a new technique is used to optimize the search process. To enhance the computational time the random walk algorithm is additionally used with the previous concept which increases the detection rate in less amount of time.

ACKNOWLEDGMENT

This research could not be finished, without the help of my faculty members and I want to express my greatest honor to them. I also want to thank all friends, who gave lots of help during the experiments.

REFERENCES

[1] R. Agrawal, J. Kiernan, R. Srikant, Y. Xu, Hippocratic databases, in: Proc. 28th Int'l Conf. on Very Large Data Bases, Hong Kong, China, 2002, pp. 143–154.
[2] Bertino, Elisa, Pierangela Samarati, and Sushil Jajodia. "An extended authorization model for relational databases", IEEE Transactions on Knowledge and Data Engineering 9.1 PP. 85-101, 1997.
[3] Kun Liu, Kamalika Das, Tyrone Grandison, Hillol Kargupta, "Privacy-preserving data analysis on graphs and social networks," In Next Generation of Data Mining, pp. 419-437, 2008.
[4] E. Zheleva, L. Getoor, "Preserving the privacy of sensitive relationships in graph data," In: Privacy, Security, and Trust in KDD, Lecture Notes in Computer Science, Vol. 4890, pp 153-171, 2008
[5] Available Online at: http://web.mit.edu/infoprotect/docs/protectingdata.pdf
[6] Kapravelos, Y. Shoshitaishvili, M. Cova, C. Kruegel, and G. Vigna, "Revolver: An automated approach to the detection of evasive web-basedmalware," in Proc. 22nd USENIX Security Symposium,2013, pp. 637–652.
[7] X. Shu and D. Yao, "Data leak detection as a service", in Proceeding 8th International Conference on Security Privacy Communication Network. 2012, pp. 222–240.
[8] Batya Kenig and Tamir Tassa "A practical approximation algorithm for optimal k anonymity", Data Mining Knowledge Discovery, Springer, 2011
[9] Sumeet Bajaj and RaduSion, "TrustedDB: A Trusted Hardware-Based Database with Privacy and Data Confidentiality", IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 3, MARCH 2014.
[10] Hua Wang, Lili Sun, Elisa Bertino, "Building access control policy model for privacy preserving and testing policy conflicting problems", Journal of Computer and System Sciences 80 (2014) 1493–1503