ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Preserving The Privacy Of Semi Structured Data Through Anonymization Technique

K.Ponnazhagu[1], Dr.T.Meyyapan[2]

[1]Research Scholar, [2]Professor
Department of Computer Science and Engineering, Alagappa University,
Karaikudi-600 003, Tamilnadu, India.
[1]ponnuvsk.92@gmail.com,[2]meyyappant@alagappauniversity.ac.in

Abstract— **Data leakage is one of the major issues in data sharing and publishing. The various number of data structure and dataset schema was proposed in periodically. In Current decades, users sharing different types of data in structured and unstructured format. Researchers expose their ideas and techniques in structured format. There are many numbers of storage and complexity issues and to avoid such problems in data manipulation, the data publishers uses the semi structured and unstructured format. Data publishers publish the data owners records for statistical and analytics purposes. Privacy conservation is necessary while publishing the data. Various security models are available for solving the privacy issues of structured dataset. This approach concentrates to find out the solution for individual privacy issues in a semi structured dataset through suppression. This technique applies in a semi structured dataset and measures the information loss and privacy increased. The proposed technique achieves the protection of individual privacy against linkage attacks and decreases the privacy gain.**

Keywords— **Anonymization, Preserving Privacy, JSON, Suppression Technique.**

I. INTRODUCTION

The disclosure of sensitive data is a major treats to the privacy of individuals and organizational data. Adversaries apply various types of attacks such as record linkage attacks, attribute linkage attacks, table linkage attacks and probability linkage attacks for extracting the sensitive information from the published data. Anonymization techniques are one the best technique for protecting the sensitive data against those attacks. Anonymization techniques are such as Generalization, suppression, data perturbation, data swapping. Researchers apply those techniques and deliver many privacy models that are suitable for the structured data set. Relational data model is a structured model that incorporates for manipulating the data and storing. Difficulties arise in data manipulation and storage management. While processing more than tera bytes amount of data which design in relational model. Nowdays, Semistructured data model decreases the complexity of storage management and increase efficiency manipulation. In addition, reduce the time complexity. There exist several approaches are compatible for relational data model only. JSON is one the semi structured data model. Many social media, public service oriented websites are releasing their information in JSON model. Researches and developers access their information through APIs. In our approach, understanding the JSON file format for manipulating the data and apply the suppression technique to protect the sensitive information. In experimental analyses, the anonymization technique is applied over JSON data file and measure the time complexity of semi structured data model. The result shows the time complexity metric value is gradually decreasing..

II. RELATED WORK

The researchers contribute the methods to parse the semi structured data and only few delivers the privacy models for semi structured schema model. A logical-form driven parsing algorithm guided by strong typing constraints and an approach based on the prototype theory [1]. Automatic classification of semi-structured data, represented by frames, without any previous knowledge about structured classes [2]. Structured data fusion (SDF) is a framework for the rapid prototyping of knowledge discovery in one or more possibly incomplete data sets. In SDF, each data set—stored as a dense, sparse or incomplete tensor—is factorized with a matrix or tensor decomposition [3]. Serge abiteb oul, isolate the essential aspects of semi- structured data and applied the query languages for semi-structured data for finding the motivation in the integration of heterogeneous data [4]. Berthier-ribeiro proposed an innovative approach to extracting semi-structured data from Web sources. The idea is to collect a couple of example objects from the user and to use this information to

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

extract new objects from new pages or texts [6]. To perform the extraction of new objects, we introduce a bottom-up extraction strategy and, through experimentation, demonstrate that it works quite effectively with distinct Web sources, even if only a few examples are provided by the user. slicing stores better data utility than generalization and utilized for membership disclosure protection. Moreover advantage of preserving is that it can maintain high-dimensional information [7]. Heuristic algorithms exploit the equivalence group monotonicity of privacy constraints and adaptive ordering techniques for efficiently checking m-privacy given a set of records [8]. The consequence of using slicing is that it can switch high dimension data. Slicing conserves better data service than generalization and also prevents participation revelation.[9] Rui chena, benjamin c proposed the work that was based on local suppression to achieve a tailored privacy model for trajectory data Anonymization [10]. A novel framework for set-valued data Anonymization by partial suppression regardless of the amount of background knowledge the attacker possesses, and can be adapted to both space-time and quality-time trade-offs in a "pay-as-you-go" approach[11].

III. Proposed Work

Semi-structured data is data that is neither raw data, nor typed data in a conventional database system. It is structured data, but it is not organized in a rational model, like a table or an object-based graph. A lot of data found on the Web can be described as semi-structured. Data integration especially makes use of semi-structured data. This data model uses for sharing the data which have large size. Especially, JSON file format is one of the semi structured data. It is a standard light weight format and data management easily. Schema of the JSON file is similar to XML schema. Its schema claims the data portability much more than relational schema. That's why the data publishers prefer this format for sharing. Illustration of JSON file format is given below.

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address":
{
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",


      "number": "     212 555-1234 FREE"
    },
    {
      "type": "office",


      "number": "     646 555-4567 FREE"
    }
  ],
  "children": [],
  "spouse": null
}
```

A valid JSON document defines by an object type or an array type. In above example, one object binds the attributes, arrays. "firstname" is an attribute and its value is "John". "PhoneNumbers" is an array. It consists of collection of attributes such as type, number. In microdata, the quasi identifiers and sensitive values defines in form of either object or an array.

A. Anonymization

There are various types of Anonymization technique such as generalization, Suppression, Data Swapping, data Perturbation, etc… Suppression technique is a familiar technique which has used in various Anonymization algorithms and models. But those models and algorithms have applied into relational model based micro data. Here, in semi structure data, suppression technique applies into quasi identifier attributes to protect the sensitive attribute values. In Relational model, DML concept helps to select the quasi identifiers for applying the preserving technique. In semi structured format, the parsing process is preliminary step to identify the quasi identifier and sensitive attribute. The following steps are parsing the JSON file

```
a) Read JSON file using JSON API
b) Parse the JSON file and Store the JSON
   Objects in array A[i] where i=1,2,3..n.
c) Initialize the array variable Quasi
   identifier Q[j] where i=1,2,3..n
d) If A[i].attribute== Q[j]
e) FoundQ[k]= Object[i].attribute.value
f) Increment j value by 1
g) Increment K value by 1
h) Reset i value as 0
i) Go to step d j=n
j) else
k) Increment i value by 1
l) Go to step d until i=n
```

After parsing process, FoundQ[k], where k=1,2,3..n, contains quasi identifier attribute values. The following steps shows the suppression technique applies into founded quasi identifier values.

- QIvalues refers to founded Quasi Identifier value
- QIvalLength refers to length of Quasi Identifier value
- AQIvalues is refers to anonymized Quasi Identifier value

```
a) Initialise the QIvalues[i] values is equal
   to founded[k] where k=1,2,3..n;
b) Find the length of OILength , OILength=
   OIvalues[i].length
```

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

c) Replace the last two character of QIvalues[i] in to '**' and stores into AQIvalues[m] where m=1,2,3..,
d) Increment m by 1
e) Increment i by 1
f) Go to step a until i=n
g) Increment k by 1
h) Go to step a unitil k=n

In below section, anonymized the postcode values of JSON file through suppression. Here the result of anonymized semi structured data shows in given below.

```
"address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-31**"

"address": {
    "streetAddress": "23 2nd Street",
    "city": "New Jersey",
    "state": "JY",
    "postalCode": "10025-45**"
"address": {
    "streetAddress": "2nd Street",
    "city": "Florida",
    "state": "FL",
    "postalCode": "45321-24**"
```

Once complete the Anonymization process, the new anonymized JSON file generated and ready to publish the data for public. Anonymization process of semistructed data hopes to encourage the reduction of time complexity while comparing to Anonymization process of relational model.

IV. Experimental Result

The proposed method is applied over the published dataset which is in JSON format. The suppression technique implements into quasi identifier attribute 'postal code. The proposed method evaluates in terms of information loss.

$$Information\ Loss(I_g) = \frac{|I_g| - 1}{|D_A|} \quad (1)$$

where; $|I_g|$ is the number of domain values that are descendants of I_g

$|D_A|$ is the number of domain values in the attribute A of I_g.

Information Loss(I_g) =0 if v_g is an original data value in the table.

TABLE I
INFORMATION LOSS OF ANONYMIZED DATASET

Dataset	Information Loss
1000 records	0.562
1500 records	0.684
2000 records	0.791

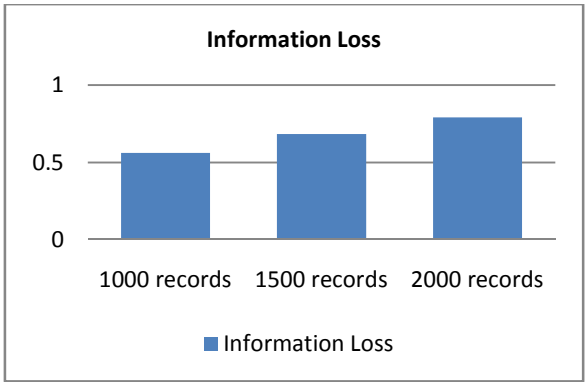

Fig1. Information loss of semistructured anonymized dataset

V. Conclusions

In Current decade, for data portability, semi structured data model being used to data sharing, Large amount of data can be managed in semi structured model. Furthermore, the privacy is one of the major worry while publishing the data. Here, the solution has given to preserve the privacy against linkage attacks in JSON format. Suppression technique has implemented in this work and believes the metric value of time complexity will be decreased that compared to Relational data model. In future, the improvement of anonymity algorithm for semi structured data model will be going to achieve the privacy in other types of semi structured data model such as BSON,YAML, etc..

References

[1] Panupong pasupat, percy liang, "Compositional semantic parsing on semi-structured tables", The association for computer linguistics, ISBN: 978-1-941643-72-3, 2015.
[2] Bernardo pereira nunes, Giseli rabello Lopes,et.all, "Automatic classification and taxonomy generation for semi-structured data", International conference on computer and information technology, Jan-2015.
[3] Sorber, Laurent, Marc Van Barel, and Lieven De Lathauwer. "Structured data fusion." Selected Topics in Signal Processing, IEEE Journal of 9.4 (2015): 586-600.
[4] Serge abiteb oul, "Querying semi-structured data", Lecture Notes in Computer Science, Database Theory — ICDT '97 -1997
[5] Jason mchugh, serge abiteb oul, "lore: a database management system for semistructured data", SIGMOD Record, 26 (3). pp. 54-66, 1997.
[6] Berthier ribeiro-netoalberto h. f. laender, "Extracting semi-structured data through examples", Eighth international conference on information and knowledge management(cikm) ,ACM,1999.
[7] Sandhya, Ch, and M. Vasavi. "Partitioning For Secure Data." SYSTEM 2.1 (2014): 1401-1404.
[8] Sujatha, Veeramreddy, and Akula Suneetha. "Collaborative Data Publishing Using Slicing And M-Privacy." IJMCA 2.5 (2014): 141-145.
[9] Reddy, Pargi Sridhar, and Ch Ravi. "A Novel Technique for Privacy Preserving Data Publishing." (2014).
[10] Chen, Rui, et al. "Privacy-preserving trajectory data publishing by local suppression." Information Sciences 231 (2013): 83-97.
[11] Jia, Xiao, et al. "ρ-uncertainty Anonymization by Partial Suppression."Database Systems for Advanced Applications. Springer International Publishing, 2014.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

From the figure 1, shows that the proposed method decreases the information loss while published thousand records. The number of publishing records is increased as well as information loss is also increased.