International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 8 54 – 58

Detection and Privacy Preservation of Sensitive Attributes Using Hybrid Approach for Privacy Preserving Record Linkage

V. Uma Rani
Associate. Prof of CSE
School of IT, JNTUH
Hyderabad (T.S), India
umarani@jntuh.ac.in

Dr. M. Sreenivasa Rao
Prof of CSE
School of IT, JNTUH
Hyderabad (T.S), India
srmeda@gmail.com

Kotra Sai Srujana
M. Tech Scholar in CS
School of IT, JNTUH
Hyderabad (T.S), India
srujna29@gmail.com

***Abstract*—** Privacy Preserving Record Linkage (PPRL) is a major area of database research which entangles in colluding huge multiple heterogeneous data sets with disjunctive or additional information about an entity while veiling its private information. This paper gives an enhanced algorithm for merging two datasets using Sorted Neighborhood Deterministic approach and an improved Preservation algorithm which makes use of automatic selection of sensitive attributes and pattern mining over dynamic queries. We guarantee strong privacy, less computational complexity and scalability and address the legitimate concerns over data security and privacy with our approach.

Keywords-*PPRL, Sorted Neighborhood Deterministic approach, Pattern mining, record linkage, data matching, entity resolution, merge-purge, privacy techniques, data matching, object identification, identity uncertainty, sensitive attribute selection.*

*****

I. INTRODUCTION

Record linkage is the process of merging the same real world data but present in a contradistinctive form as they are captured from heterogeneous sources, for the identification of matching record pairs, mining and further analyzing the data. This has a long tradition in both the statistical and the computer science literature.

In this data-driven world, many organizations spawn and store large volumes of data. So, when record linkage is applied within a single organization generally privacy and confidentiality are not of great concern (assuming there are no internal threats). However, when data are linked from several organizations, privacy and confidentiality considerations are vital.

For instance, if the personal information about customers is used in the linking of databases across organizations, then the privacy of this information must be preserved. Individual databases can contain highly sensitive data, such as their medical or financial details.

There is a chance of sensitive data leakage when linked, such as personal information in case of individuals, confidential information of an organization etc...

But this integration of data resources and data mining are important in identifying and studying the data for making strategic decisions which would benefit the organization. For example, a research group is interested in analyzing the effects of car accidents upon the health system, most common types of injuries, financial burden upon the public health system, and general health of people after they were involved in a serious car accident. They need access to data from hospitals, doctors, and Car and health insurers and from the police.

Data Integration from heterogeneous sources is not an easy task because the synonyms used are different and unique identifier does not exists in linking them. Privacy Preserving Record Linkage (PPRL) is the problem where the only extra knowledge that each source gains relate to the records which are shared among the participating sources after the data integration ensuring the privacy of sensitive personal data

II. RELATED WORK

Record matching (or linkage) is a rather old yet important area of research. There are plentiful methods that have been proposed to address the problem as record matching is an old yet important area of research. A detailed analysis of all major currently used methods can be found in [1]. Approximate string matching methods compare strings to possible typographical errors. These methods fall into three major categories: Token-based methods, distance-based methods and phonetics based methods.

Token-based methods calculate tokens of the strings to be matched and then count the number of common tokens. N-grams based methods fall into this category [2]. Distance-based methods calculate the distance between the strings. Some of the most widely used methods are Levenshtein distance, the Jaro and Jaro-Winkler metrics. Conversely, phonetics based methods make use of certain string transformations to take advantage of the way words sound for purging the effect of various typing and spelling errors. Typical examples of this class include Soundex [4], Metaphone [5], ONCA [6], and NYSIIS [7].

Bloom filter-based PPRL is also a frequently used technique. The method uses encrypted personal identifying information (bloom filters) in a probability-based linkage framework [12].

III. BACKGROUND

The background study here portrays few algorithms that are required to represent our methodology. We here illustrate a running example and is used throughout the paper. More specifically, we emphasize the phonetic algorithms and distance-based matching methods. We also present the operation of matching algorithm, an extension of Sorted Neighborhood approach.

A. Merging Algorithms

1. *Phonetic Algorithms*

Phonetic algorithm is a calculation to coordinate and match words based on their pronunciation. Phonetic calculations have been comprehensively utilized as a part of the past for record coordinating performed on names. The primary element of the phonetic algorithms is their adaptation to fault tolerance against typographical mistakes. For delineation purposes, we will utilize Soundex [9] in this paper. However, our procedure can be effectively connected to other phonetic calculations. The operation of Soundex is very clear: for each word to be encoded certain rules of grouping similar sounds are applied. The outcome is a four character hash that represents the pronunciation of the word. This hash comprises of a capital letter taken after by three digits. For instance for the word "Desmet", its Soundex code is D253.

2. *Distance based Methods*

Distance-based methods employ functions that map a pair of strings to a real number [9].
Levenshtein distance [10] is the best known representative of distance functions. It quantifies the minimum number of operations required (insert, delete, replace) to transform one string to another. Here, two strings are said to match if their distance is not as much as d operations, d> 0.

B. Blocking Algorithms

1. *Re-sampling Strategy*

A greedy re-sampling heuristic based on Sparse Map is used to map values into a vector space at lower computational costs. However, the experimental results presented by Scannapieco etal (2007) indicates that the linkage quality is influenced by the greedy heuristic re-sampling strategy.

2. *Blend of anonymization and cryptographic techniques*

A hybrid approach that joins anonymization techniques and cryptographic techniques to tackle the private record linkage problem is proposed by Inan et al. (2008). This method uses value generalization hierarchies in the blocking step, and the record pairs that cannot be blocked are compared in a computationally expensive secure multiparty computation (SMC) step using cryptographic techniques.

3. *Encoded Phonetic Codes*

Using the one-to- many property of phonetic codes,an approach is proposed by Karakasidis & Verykios (2009) for performing approximate matching in PPRL. The attribute values are encoded using a phonetic encoding algorithm such as Soundex (Christen 2006a) and the resulting phonetic codes are mixed with randomly generated phonetic codes and sent to a third party to perform matching. The approach is secure and efficient for approximate matching but is not appropriate for linking records based on numerical attributes, since phonetic codes are not suitable for numerical values.

IV. OUR NEW APPROACH

Our work presents two algorithms:

First, *the Merging algorithm* aims at high performance PPRL. We modify the well-known Sorted Neighborhood algorithm over the standardized data so that it operates on all types of data.

Second, the *Blocking algorithm* aims to conceal the sensitive information of the individual using pattern mining over dynamic queries.

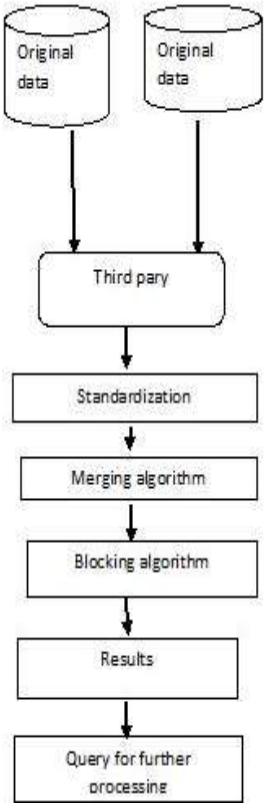

Figure 1: Block diagram of PPRL algorithm

*A. **Merging Algorithms***

Many classical techniques were proposed by different authors to tackle the private record linkage problem, which differ in computational costs, efficiency, in privacy notions, scalability etc. In this study, we divide our Merging algorithm into the following three steps:

- *Standardization using secure transformations, Secure Multiparty Computation [4], and*
- *Matching using sorted neighborhood Deterministic method [].*

1. *Standardization using secure transformation*

Secure transformation techniques aim to perform the linkage of the records after few transformations have been applied to the original data.

For instance, the distinctive formatting styles of records look different but all refer to the same entity with the same logical identifier values. Record linkage strategies would bring about more accurate linkage if these values were first standardized into a consistent format (e.g., all names are "address formats, name formats like first name ,middle name, last name, all dates are "YYYY/MM/DD"). Standardization can be accomplished through simple rule-based data transformations.

DataSet	*Name*	*Car Info*	*Insurance Company*
Data Set 1	James Arnold	Ford Figo	New India Assurance, Kolkata, India
Data Set 2	Arnold James	Figo Ford- Xsi	New India Assurance Co. Ltd, Kolkata, India
Data Set 3	A. James	Ford xsi	NIACL, Kolkata, India

TABLE I: Sample data to be standardized

2. *Secure Multiparty Computation*

The typical scenario involves three parties, where two parties have the data, and using secure transformation techniques, the data are sent to a third party, whose task is to perform the matching using Sorted Neighborhood Deterministic record linkage . The data are matched in such a way that the sensitive information (like SSN, name and other personal details of customer) are hidden from the third party.

3. *Sorted Neighborhood Deterministic method*

Once the data are standardized, record linkage, called deterministic or rules-based record linkage generates links based on the number of individual identifiers that match among the available data sets. [2].The following matching algorithm is used.

- Identify the source dataset
- Populate columns from the data dictionary as per the source table
- Identify the target dataset
- Populate columns from the data dictionary as per the target table
- Identify a matching attribute one each from the source and the target
- Identify the merged dataset.
- Choose the blocking attribute – Sensitive data from the merged dataset.
- Prepare a dynamic SQL that comprises all the selected cols in the merged dataset with data type varchar
- Drop any previously created merged dataset
- Construct the Create table statement dynamically with the selected columns and execute to create the table.
- Fetch values for all the matching columns from both the tables and insert into merged dataset.
- Save the blocking attribute with table name into privacy dataset.

Data Set		*SSN*	*Name*	*Car Info*	*Insurance Company*	*ZIP*
Set A	1	325466776	Joy, Cassandra	Honda city	ICICI	763444
	2	325466776	Joy, Cassandra	Honda City	ICICI	763456
	3	764356757	Robert, williams	Ford figo	BAJAJ	763444
	4	876635425	Mary, smith	Tata tiago	NICL	376543
Set B	1	764356757	Robert, JK	Ford figo		
	2		Joy, Bill	Renault Duster	NICL	763444

TABLE II: Sample dataset to be merged

Algorithm for the detection of blocking attributes:

The two data sets are merged as a sequence of one single dataset using Sorted Neighborhood, all the records in this data set are sorted with based on one attribute called RBL approach a window of size *w* is set over this merged dataset, the first record is matched with the rest of the records in the window. Two records are said to match via a deterministic record linkage procedure if all or some identifiers (above a certain threshold) are identical. All the matching records if any are there in the window are copied to another dataset. Then the window is slided to next w records, until there are no records for the window, this is repeated.

When the entities in the data sets are identified by a common identifier, or when there are several representative identifiers (e.g., name, address, ZIP etc… when identifying a person) whose quality of data is relatively high, Deterministic record linkage is the most preferred option.

As an example, consider two standardized data sets, Set A and Set B, that contain different bits of information about customers of Car Insurance companies. A variety of identifiers: Social Security Number (SSN), name and ZIP code (ZIP).

The most simple deterministic record linkage strategy would be to pick a single identifier which can be uniquely identified, say SSN, and declare that records sharing the same value identify the same person while records not sharing the same value identify different people. In this example,

deterministic linkage based on SSN would create entities based on A1 and A2; A3 and B1; and A4. While A1, A2, and B2 appear to represent the same entity, B2 would not be included into the match because it is missing a value for SSN.

Missing identifiers involves the creation of additional record linkage rules. One such rule in the case of missing SSN might be to compare name, insurance company, and ZIP code with other records in hopes of finding a match. In the above example, this rule would still not match A1/A2 with B2 because the names are still slightly different: standardization put the names into the proper (Surname, Given name) format but could not discern "andra" as a nickname for "cassandra". Running names through a phonetic algorithm such as Soundex, NYSIIS, or metaphone, can help to resolve these types of problems.

A. *Blocking Algorithm*

This hides the sensitive information of the individual using pattern mining over dynamic queries.

The user has to pick up the Blocking attributes over the merged data and these blocking attribute details and merged dataset name are saved in another privacy table for which the permissions are denied for every other user.

Pattern mining over dynamic queries

Any third party are allowed to query dynamically over the merged dataset, the following Blocking algorithm is used to hide the sensitive attributes.

- Prompt query to retrieve
- Divide the query into two types
  a) Where all the columns are retrieved by using the operator '*'
  b) Where specific columns are retrieved by specifying column names delimited by ",".
- Tokenize the query to identify the table name and column parameters.
- Check for table existence in the privacy table and determine the blocking attribute.
- if (qry.IndexOf("*") >= 0) then
  - retrieve all the columns from the data dictionary
  - Reconstruct the query fetching all the columns from the specified table except for the blocking attribute.
  - Execute the query to display values in the grid except the blocking attribute
- Else
  - String fs=battr + ",";
- if (qry.IndexOf(fs) >= 0)
  - qry = qry.Replace(fs, " ");
- else{
  - fs = "," + battr;
  - qry = qry.Replace(fs, " ");

  }
- Construct the SQL ignoring the blocking attribute and execute.
- Display aligned dataset.

Selection of Blocking Attribute:

We determine the blocking attribute by the following steps.

Step1: Compute the Attribute evaluation Measures
Step2: Rank the Attributes based on Attribute evaluation
. measures
Step3: Compute Relative Sensitivity Ranks of attributes
Step4: Compute absolute Sensitivity Ranks of attributes
Step5: Select the attributes whose rank is below the Threshold Value as sensitive attributes for blocking.

All sensitive attributes are of two types:

- Identity Related Information,
- Confidential Information.

We are finding blocking attributes using attribute measures like Information Gain, Relief, CFS (Correlated based Feature Selection), Gain Ratio, Correlation Attribute Evaluation, One R Attribute Evaluation, Symmetric Uncertainty Attribute Evaluation, Symmetrical Uncertainty Attribute Evaluation, Wrapper Subset Evaluation, Principal Component Analysis.

V. EVALUATION

Here, we provide detailed analysis of operations taking place at both the Blocking and the Matching Components. The evaluation is made in terms of efficiency and complexity and protocol security.

A. Efficiency and Complexity

The phonetic codes do not offer detailed matching, so there are an increased number of mismatches, having simultaneously increased sensitivity to specific alterations. This renders them unsuitable for detailed matching evaluation. Our novel proposal is a hybrid approach which uses Deterministic matching with a Sorted Neighborhood approach and is efficient and less complex, since it reduces the matching space. Sorting each field of RBL requires O(n log n) and scanning O(n) operations, reducing the candidate pairs significantly. Comparing all by all matching fields would require $O(n^2)$ comparisons. The decreased complexity of our approach allows applying the blocking passes more than once with different blocking keys.

B. Privacy Analysis

We here present an analysis that focuses on two aspects, the information gained by each of the data holders and the information gained by a possible eavesdropper over the transmission channel, to evaluate the privacy offered to the integrated data by our protocol. Private data belonging either to the matching or to the blocking dataset are saved in another dataset for which no privileges are given to the end user and the data in the privacy table are encoded using secure hash function with an encrypted key. Therefore, the attacker should be aware of the key used in the hash function. Therefore, the attacker should pose a brute force attack to identify the hashing key used and the type of matching algorithm used, since all data are broken into tokens depending on the agreed matching technique.

VI. EXPECTED RESULTS

A lot of study and analysis have been made on the present method and huge computations have been applied on large number of data sets with in different environments. A comparative analysis is made between the present method and several previous methods in a well efficient manner and also shown in the below figure in the form of the graphical representation and is explained in an elaborative fashion respectively. There is a huge challenge for the present method where accurate analysis is made, where the major aspect is the data matching based on different hybrid approaches and data preservation aspects using pattern mining in a well effective manner and also analysis of the sentiment based strategy relative to the positives followed by the negative in an accurate fashion respectively. Here we finally conclude that the present method is effective, scalable and efficient, in terms of the analysis based aspect which is related to the performance based strategy followed by the accurate outcome of the entire system in a well oriented fashion respectively.

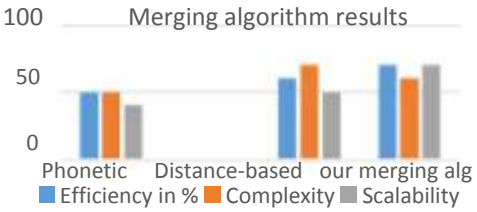

Figure 2: Graph for Merging Results

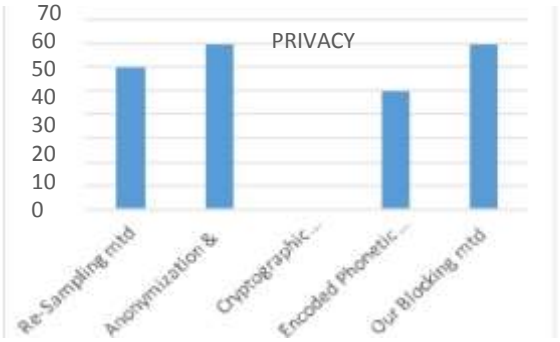

Figure 3: Graph for blocking results

VII. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a novel method for automatic detection of sensitive attributes and privacy preserving blocking. We have proved that our approach is secure, less complex, fast, accurate and robust and exhibits better behavior than state-of-the-art methods. Our next steps include more extensive experimentation, in order to assess scalability, with different ranking functions and real world data.

Moreover we would like to develop a faster yet secure PPM method for numeric fields. Finally, we aim at developing a method for PPM which, as the privacy preserving blocking method we have presented, will operate independently at each site and will be suitable for any type of data field.

REFERENCES

[1]. A. K. Elmagarmid, P. G. Ipeirotis, and V.S. Verykios, "Duplicate record detection: a survey," IEEE Transactions on Knowledgeand Data Engineering, vol. 19, no. 1, pp. 1-16, 2007

[2]. L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, L.Pietarinen, and D. Srivastava, "Using q-grams in a dbms for approximate string processing," IEEE Data Engineering Bulletin, vol. 24, no. 4, pp. 28-34, 2001.

[3]. M. K. Odell and R. C. Russell, US Patent Number 1261167, 1918.

[4]. A Sorted Neighborhood Approach to Multidimensional Privacy Preserving Blocking Alexandros Karakasidis and Vassilios S. Verykios *School of Science and Technology Hellenic Open UniversityPatras, Greece*

[5]. L. Philips, "Hanging on the metaphone," Computer Language, vol. 7, no. 12, pp. 39-43, Dec. 1990.

[6]. L. E. Gill, "OX-LINK: the Oxford medical record linkage system," Record Linkage Techniques--1997: Proceedings of an International Workshop and Exposition, Arlington, VA, 1997, pp.15-33.

[7]. R. L. Taft, Name Search Techniques. Special Report / New York State Identification and Intelligence System, Albany, NY: Bureau of Systems Development, 1970.

[8]. W.Cohen P. Ravi Kumar, and S.E. Fienberg, "A comparison of String Distance metrics for name-matching tasks," Proceedings of the IJCAI 2003 Workshop on Information Integration on the wen, Acapulco,Mexico,2003,pp. 73-78.

[9]. V. Levenshtein, "Binary Codes capable of correcting deletions, insertions and reversals," soviet Physics Doklady, vol. 10, no.8, pp.707-710,1966.

[10]. R.Schnell, T. Bachteler, and J.Reiher, "Privacy-Preserving record linkage using bloom fillers," *BMC Medical Informatics and Decision Making*, vol 9, no. 1, pp. 44+, August 2009.

[11]. Dinusha Vatsalan. Vassilios S. Verykios. "a Taxonomy of Privacy-preserving record linkage techniques," *information Systems*, vol 38,issue 6, September 2013.

[12]. Roos, LL; Wajda A (April 1991). "Record linkage strategies. Part I: Estimating information and evaluating approaches.". *Methods of Information in Medicine*30 (2):117–123. PMID 1857246. Retrieved 11 November 2011.