www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 4 April, 2014 Page No. 5507-5513

Privacy Preservation of Sensitive Attributes Using Hybrid Approach

M. Geetha *1, V. Uma Rani 2

1. Asst. Prof, Dept of MCA, Loyola Academy Degree and PG College, Old Alwal, Secunderabad-10, India, part time MTech(CS) student of SIT, JNTUH.
2. Asst. Prof, Dept of CSE, SIT, JNTUH, Hyderabad (A.P), India.

ankerlageetha@yahoo.com

ABSTRACT

Privacy Preserving Record Linkage (PPRL) is widely used in data mining applications which aims to integrate data from different heterogeneous data sources while hiding the private information. In this paperwe propose a new algorithm for merging two datasets using Sorted Neighborhood Deterministic approach and a new Preservation algorithm that uses Pattern mining over dynamic queries. In contrast to the existing techniques our approach guarantees strong privacy less computational complexity and is scalable over large datasets. We provide empirical evidences to prove that our method is secure, fast and efficient than the existing methods.

Keywords: PPRL, Sorted Neighborhood Deterministic approach, Pattern Mining

1. INTRODUCTION

Record linkage involves merging the data from different multiple data sources using data integration and data mining tasks to identify the records that refer to the same real world entity.

Many organizations these days generate or collect large volumes of data, for example medical data of different patients in different hospitals, history of the customers who borrow loans, details of the customers who book the tickets for flights etc. If we are able to interconnect these data and analyze them it would be beneficial to the organizations.

For example if we are able to interconnect the data from a health organizations and the data of the customer who are going by flight it could prevent the customers infected by others, and not only these, the data collected from different hospitals helps us to analyze the different treatments for different ailments.

To integrate data from different sources is not an easy task, since the data are stored by different organization using different synonyms, and there is no unique identifier to link these records.

Privacy Preserving Record Linkage (PPRL) is the problem where data is integrated in such a way that after the integration process, the only extra knowledge that each source gains relates to the records which are shared among the participating sources. i.e., the personal data of the user is not disclosed.

2. RELATED WORK

Record matching (or linkage) is a rather old yet important area of research. As such, numerous methods have been proposed to address the problem. A detailed analysis of all major currently used methods can be found in [1]. Approximate string matching methods consider comparing

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 measure the differences between 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].

3. BACKGROUND

In this section, we provide the necessary background requiredto present our methodology, along with a running example usedthroughout the rest of our paper. Specifically, we describe thephonetic algorithms and distance-based matching methods.Moreover, we present the operation of matching algorithm which is an extension of Sorted Neighborhood approach.

3.1 MERGING ALGORITHMS

A. Phonetic Algorithms

Phonetic algorithm is an algorithm to match words based on their pronunciation. Phonetic algorithms have been broadly used in the past for record matching performed on names. The main feature of the phonetic algorithms is their fault tolerance against typographical errors. For illustration purposes, we will use Soundex [9] in this paper. However, our methodology can
be easily applied to other phonetic algorithms. The operation of Soundex is quite straightforward: for eachword to be encoded certain rules of grouping similar sounds are applied. The result is a four character hash that represents the pronunciation of the word. This hash consists of a capital letter followed by three digits. For example for the word “Cooper”, its
Soundex code is C160.

B. Distance-Based Methods

Distance-based methods employ functions that map a pair of strings to a real number [9] Levenshtein distance [10] is the bestknown representative of distance functions. It measures 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 less than d operations, d> 0.

3.2. BLOCKING ALGORITHMS

A. Re-Sampling method

A greedy re-sampling heuristic based on SparseMap is used to map values into a vector space at lower computational costs. However, the experimental results presented by Scannapieco et al. (2007) indicate that the linkage quality is affected by the greedy heuristic re-sampling method.

B. Combination of anonymization and cryptographic techniques

A hybrid approach that combines anonymization techniques and cryptographic techniques to solve 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.

C. 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 2006*a*) 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 suitable for linking records based on numerical attributes, since phonetic codes are not suitable for numerical values.

4. OUR NEW APPROACH

Our work introduces two algorithms
First, *Merging algorithm* that aims at high performance PPRL. We modify thewell-known Sorted Neighborhood algorithm over the standardized data so that it operates on all types of data.
Second, *Blocking algorithm* which aims to hide the sensitive information of the individual using pattern mining over dynamic queries.

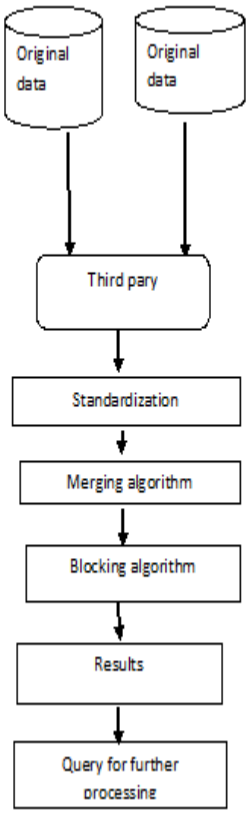

Fig 1: Block diagram of PPRL algorithm

4.1. Merging Algorithm

There are already many classical techniques proposed by different authorsto tackle the private record linkage problem, which differ in computation cost, efficiency, in privacy notions, scalability etc.
In our paper we divide our Merging algorithm into the following three steps:

Standardization using secure transformations,

SecureMultiparty Computation [4], and Matching using sorted neighborhood Deterministicmethod [].

4.1.1. Standardization using secure transformation

Secure transformation techniques aim to perform the linkageof the records after some transformations have been applied to theoriginal data.

For example, the different formatting styles of records look different but all refer to the same entity with the same logical identifier values. Record linkage strategies would result in more accurate linkage if these values were first *standardized* into a consistent format (e.g., all names are "Surname, Given name", all dates are "YYYY/MM/DD", and all cities are "Name, 2-letter state abbreviation"). Standardization can be accomplished through simple rule-based data transformations.

Data set	Name	Date of birth	City of residence
Data set 1	William J. Smith	1/2/73	Berkeley, California
Data set 2	Smith, W. J.	1973.1.2	Berkeley, CA
Data set 3	Bill Smith	Jan 2, 1973	Berkeley, Calif.

Tab 1:sample data to be standardized

4.1.2Secure Multiparty Computation

The typical scenario involves three parties, wheretwo parties have the data, and using secure transformation techniques, they send the data to a third party whose task is to performthe matching using Sorted Neighborhood Deterministic record linkage the data is matched in such a way that the sensitive information (like name and other personal details of the patient who is suffering from cancer etc.) is hidden from the third party.

4.1.3 Sorted Neighborhood Deterministic method

Once the data is 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 both

Data Set	#	SSN	Name	DOB	Sex	ZIP
Set A	1	000956723	Smith, William	1973/01/02	Male	94701
	2	000956723	Smith, William	1973/01/02	Male	94703
	3	000005555	Jones, Robert	1942/08/14	Male	94701
	4	123001234	Sue, Mary	1972/11/19	Female	94109
Set B	1	000005555	Jones, Bob	1942/08/14		
	2		Smith, Bill	1973/01/02	Male	94701

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 columnss 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.

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, this is repeated until there are no records for the window.

Deterministic record linkage is a good option when the entities in the data sets are identified by a common identifier, or when there are several representative identifiers (e.g., name, date of birth, and sex when identifying a person) whose quality of data is relatively high.

As an example, consider two standardized data sets, Set A and Set B, that contain different bits of information about patients in a hospital system. The two data sets identify patients using a

Tab 2: sample datasets to be merged

variety of identifiers: Social Security Number (SSN), name, date of birth (DOB), sex, and ZIP code (ZIP). The records in two data sets (identified by the "#" column) are shown below:
The most simple deterministic record linkage strategy would be to pick a single identifier that is assumed to be uniquely identifying, 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, date of birth, sex, 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 "Bill" as a nickname for "William". Running names through a phonetic algorithm such as Soundex, NYSIIS, or metaphone, can help to resolve these types of problems.

4.1.2 BLOCKING ALGORITHM

Aims to hide the sensitive information of the individual using pattern mining over dynamic queries.
Over the merged data the user is asked to pick up the Blocking attributes, 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
    - Where all the columns are retrieved by using the operator '*'
    - 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.

5. EVALUATION

In this section, we will provide detailed analysis regarding operations taking place at both the Blocking and the Matching Component. The evaluation is made in terms of efficiency and complexity and in terms of protocol security.

A. Efficiency and Complexity

The phonetic codes do not offer detailed matching, leading to increased number of mismatches, having simultaneously increased sensitivity to specificalterations. This renders them unsuitable for detailed matching evaluation. Our novel proposal is a hybrid approach which uses Deterministic matching with a Sorted Neighborhood approach isefficient and less complex, since it reduces the matching space. Sorting each field of RBLrequires O(nlogn) and scanning O(n) operations, reducing the candidate pairs significantly. Comparing allby all matching fields would require O(n2) comparisons. Thedecreased complexity of our approach allows applying theblocking passes more than once with different blocking keys.

B. Privacy Analysis

We will present an analysis focusing 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 eitherto 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 shouldpose a brute force attack to identify the hashing key used andthe type of matching algorithm used, since all data are brokeninto tokens depending on the agreed matching technique.

6. EXPECTED RESULTS

A lot of analysis has 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 to that of the several previous methods in a well efficient fashion 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, oriented 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.

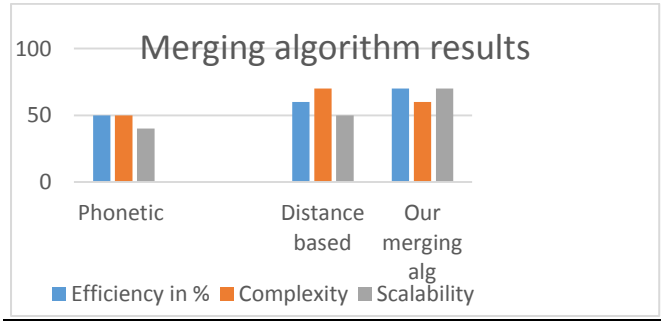

Fig 2: Graph for Merging Results

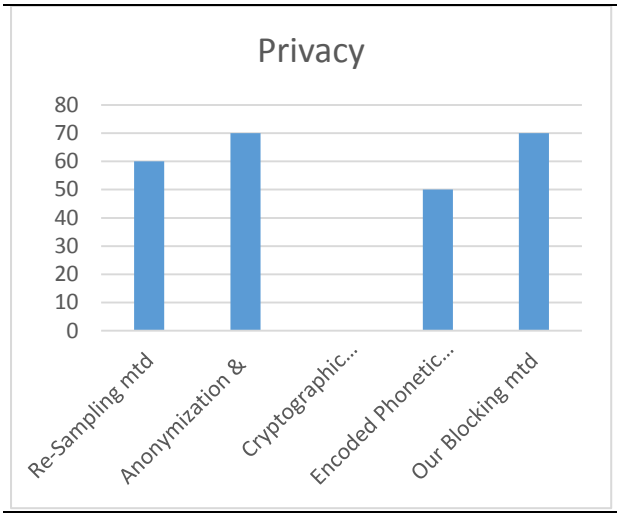

Fig 3: Graph for Blocking results

7. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a novel method for 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 wish 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 Sorted Neighborhood Approach to Multidimensional Privacy Preserving Blocking Alexandros Karakasidis and Vassilios S. Verykios*School of Science and TechnologyHellenic Open UniversityPatras, Greece*

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

[3]. 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.

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

[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]. 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.

[9]. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," Proceedings ofthe IJCAI 2003 Workshop on Information Integration on the Web, Acapulco, Mexico, 2003, pp. 73-78.

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

[11]. R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preservingrecord linkage using bloom filters," *BMC Medical Informaticsand Decision Making*, vol. 9, no. 1, pp. 41+, August 2009