IJCSN International Journal of Computer Science and Network, Volume 4, Issue 6, December 2015
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 0.417

A Survey on Privacy of Record Linkage by using Locality Sensitive Hashing Method and Secure Multi-Party Computation

[1] **Kirti Satpute,** [2] **P. B. Mali**

[1,2] Department of Computer Engineering,
Smt.Kashibai Navale College Of Engineering
Savitribai Phule, Pune University
Pune, 411041, Maharashtra, India.

Abstract - The main motivation behind this paper is to present a vital overview of Locality-Sensitive Hashing which is base blocking method for preserving the security of record linkage. The proposed LSH method is used for guess the candidate record pairs, which goes through an anonymous transformation. In this survey paper Cosine based approach is proposed to compare the formulated record pair's in homomorphism way.

Keywords - *Locality-Sensitive Hashing, Blocking, Jaccard, Record Linkage.*

1. Introduction

The entity of Real-world is present by the distinct set of features in distinct data sets. This different representation of entities becomes challenging issues in form of identical and connecting records with respect to the similar original entity across distributed data sets. If in case the sets of data contain private data, then problem becomes even harder due to privacy concerns.

This method is for recognizing and involving distinct representation of the similar real-world things. Overall numerous information sources is termed as the record linkage issues. Record Linkage characterized into two steps. The initial step is about potential searching of matched pairs and finally performing actual matching of pairs which is implemented in an exact manner otherwise in an approximate manner. An exact matching of both records can be executed based on binary decision problem with two possible outcomes such as the match or mismatch of these records. The anonymization of the recorded information is performed in order so that private information in a record is not disclosed to third parties other than the owner. The anonymization technique is time and cost efficient.

Locality-Sensitive Hashing technique is efficient and widely-used method for matching of similar items among huge data sets.

Secure Multiparty Computation is the one of the subfield of cryptography for creation of methods for computing a function of inputs by keeping inputs as private and also deals with to solve privacy preserving data mining problems.SMC distributes the tasks in secure manner

When we share private information we need to take care of privacy preserving technique in order to consider privacy concerns. There are two methods for privacy preserving are cryptographic and sanitization.

Sanitization techniques mainly include privacy metrics that determine the huge of privacy preservation. Cryptographic method does not provide the accuracy to achieving privacy. These algorithms provides functional to personal data which are transformed to sequence of functions. After that by utilizing SMC technique precise results are obtained.

2. Motivation

Big quantity of data which are being composed from both by organizations in the private and public regions as well as by individuals this data from various outcomes necessary to be incorporated and linked. As and when databases are linked across organizations, maintaining confidentiality and privacy is very important.

3. Literature Survey

In paper [3] to accomplish secure blocking several schemes and security for record blocking. The issues are to rapidly matching records such as record linkage problem from two independent sources without disclosing privacy to the other parties is considered. Paper mainly focuses on secure blocking scheme. The blocking schemes described are classified by varying privacy as per less data sharing at the outlay of ability. Scheme called Record-aware solves the issues of simple blocking by pairing an identifier with each hash signature. In future the ability of presented blocking schemes is further enhanced by introducing a second phase of blocking: Jaccard.

In paper [4] to accomplish linkage of record without disclosing anything about the records which are non linked is the main problem of record linkage of privacy preserving. To perform privacy preservation of record linkage for string records this paper presented a novel length of variable frequent length embedding strategy on grams based . To build the embedding base, frequent variable length grams are used which are extract from the original database based on the privacy of differential In future work, author plan to enhance the allocation strategies for the prefix tree miner and the threshold schemes for matching in embedded space. The benefit of framework of proposed work is to provide formal, verifiable privacy preservation guaranty and accomplishes superior scalability.

In paper [7] to accomplish a scalable adaptive scheme for clustering or matching entity names that they can be trained to enhance performance in a particular domain. Adaptive matching and clustering is motivated by the model of "learning to order" and this scheme considers adaptive ordering systems. The issues of learning to order is address and solved by supervised learning of a binary ordering relation, followed by a greedy method for construction of a total order specified a group of binary ordering decisions. Results with this method are comparable to or better than results achieved by clustering or matching with two plausible fixed distance metrics.

In paper [5] introduces locality sensitive hashing scheme for the matching search in high-dimensional spaces and for the estimated Nearest Neighbor issues, which is depend on p-stable distributions. If data fulfilling certain "bounded growth" condition then this algorithm searches the accurate nearest neighbor. As compare to previous scheme, LSH scheme directly works points of Euclidean space with no embeddings. As a result, the resultant query time leap is complimentary of big factor and easier for accomplish. Due to this approach presented data structure is up to 40 times faster than kd-tree. Proposed algorithm provides advantage such as easier and simplest for implementation.

In paper [6] novel and efficient approach to resolve the record-linkage issues described. The issue of record-linkage occurs obviously in the data cleansing that typically precedes data mining and analysis. For each and every attribute of records, author first map values to a multidimensional Euclidean space that conserve domain-specific similarity. The selected attribute is utilized for determining same pairs of records using the multidimensional similarity join. The benefits of proposed algorithm provide very good effectiveness and accuracy and also this approach is very extendable, since various techniques are used.

In paper [2] introduces method of statically informed generate to bloom filter encodings that incorporates bits from multiple fields. Tradeoff between security and accuracy is enables using a user-specified proposed method. New encodings are avoiding a cryptanalysis attack that successful against existing methods of FBF encodings. This approach provides the foundation for a private record linkage system.

In paper [8] author presented cryptographic Privacy Enhancing Technologies for biometric face recognition systems. This procedure is a proficient procedure for securely evaluating two Pailler encrypted numbers. Author provided a proficient procedure which allows matching an encrypted image viewing a face against a database of facial templates in the biometric itself although the detection result is secreted from the server which performs the matching.

In paper [10] to accomplish Cryptographic Long-term Key. It comprises of only one bloom filter into which identifiers are in this manner stored. Tests on reproduced databases yield linkage results comparable to non encoded identifiers and better than results from up to this point existing techniques.

Table 1: Survey Table

Sr. No	Paper	Techniques and methods used	Findings
1	Blocking Aware Private Record Linkage	Pairing an identifier with every hash signature.	Perform large-scale record linkage without revealing privacy. resulting in heavy computational costs
2	Frequent grams based Embedding for Privacy Preserving Record Linkage [4]	String records, novel frequent variable length grams based embeddin g strategy.	Provides formal, provable privacy guarantees and achieves better scalability. The allocation strategies need be enhanced.
3	Learning to Match and Cluster Large High Dimensional Data Sets For Data Integration [5]	a scalable adaptive scheme for clustering	Results with this with two plausible fixed distance metrics. Results are not persistent
4	Locality-Sensitive Hashing Scheme Based on p-Stable Distributions [6]	New Locality Sensitive Hashing scheme.	This approach presented data structure is up to 40 times faster than kd-tree. LSH scheme works directly on points in the Euclidean space without embeddings

4. Proposed Work

We proposed a Cosine based approach in order to compare the formulated record pair's homomorphically. Proposed method provides accuracy in the less possible running time as compare to previous approach. In proposed system architecture [1] records are taken as input then record is annonymize. Using SMC cosine locality sensitive hashing. We generate record linkage as output.

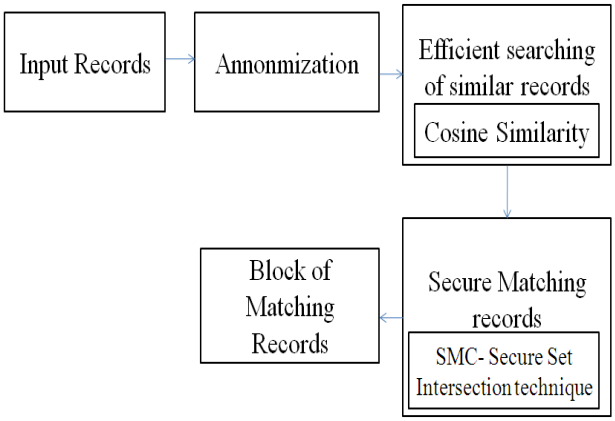

Figure 1: System architecture

5. Algorithm Used

1. Initialize Database with records
2. Anonymize the records compare two records
   If (common record)
   Then {give an id to common record
   Store the common record}
3. Search for similar record
   Apply (locality sensitive hashing tech)
4. Secure matching of record pairs
   Get (common block suing id)
   {
   Compute (distance matric)
   }
   SML (distance)
   {
   Generate for annonymize record
   Encrypt (Key + ID of record)
   Divide ID pair into subset (i.e.SA and SB)
   Encrypt the sets
   Matched pair added to M (set)
   }

6. Mathematical Modeling

6.1 Input

6.1.1 Records

$I = \{i_1, i_2, i_3 \ldots in\}$

Where, I is the input record set and $i_1, i_2, i_3, \ldots..,i_n$ are the distinct number of records.

6.2 Process

6.2.1 Anonymization of Records

A= {a_1, a_2, a_3… a_n}

Where, A is the set of anonymization record and a_1, a_2, a_3… a_n are the distinct number of annonymize records.[11]

6.2.2 Efficient Searching Of Similar Records

S= {J, E, H}

J= Jaccard LSH
E= Euclidian LSH
H= Hamming LSH
Similarity Coefficient (X, Y)

$$\textit{Cosine Coefficient} = \frac{|X \cap Y|}{|X|^{1/2} \cdot |Y|^{1/2}} \quad \ldots\ldots\ldots\ldots\ldots (1)$$

$$\textit{Jaccard Coefficient} = \frac{X \cap Y}{|X| + |Y| - |X \cap Y|} \quad \ldots\ldots\ldots\ldots (2)$$

6.3 Output

6.3.1 Securing Matching of Records

M= {m_1, m_2, m_3.… m_n}

Where, M is the set of Securing Matching of Record and m_1, m_2, m_3.... m_n are the number of matching record.

7. Conclusion

Linking large collections of records as well as protecting their privacy has become issues in the foundation of the domain of privacy preserving record Linkage. This paper surveys various techniques of privacy preserving record linkage and discusses their advantages and limitations and compares them. Anonymity requirement is the majority important parameter to alter the quantity protection and disclosure risk.

We gives an Cosine based approach in turn to compare the prepared record pair's homomorphically. In future we will focus on record pairs that have a higher distance than the defined threshold which results to low Pairs Quality rates. So the further research work on this issue is necessary in turn to filter further the contents of the blocks and discard those pairs from the identical step.

References

[1] Dimitrios Karapiperis and Vassilios S. Verykios, Member, IEEE, "An LSH-Based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage" ieee transaction on knowledge and data engineering, vol. 27, no. 4, april 2015.

[2] E. Durham, M. Kantarcioglu, Y. Xue, C. Toth, M. Kuzu, and B. Malin, "Composite Bloom filters for secure record linkage," IEEE Trans. Knowl. Data Eng., vol. PrePrints, no. 99, 2013.

[3] A. Al-Lawati, D. Lee, and P. McDaniel, "Blocking-aware private record linkage," in Proc. 2nd Int. Workshop Inf. Quality Inf. Syst., 2005, pp. 59–68.

[4] L. Bonomi, L. Xiong, R. Chen, and B. C. M. Fung, "Frequent grams based embedding for privacy preserving record linkage," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 1597–1601.

[5] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni, Locality-sensitive hashing scheme based on p-stable distributions," in Proc. 20th Symp. Comput. Geometry, 2004, pp. 253–262.

[6] L. Jin, C. Li, and S. Mehrotra, "Efficient record linkage in large datasets," in Proc. 8th Int. Conf. Database Syst. Adv. Appl., 2003, pp. 137–146

[7] W. Cohen and J. Richman, "Learning to match and cluster large high-dimensional datasets for data integration," in Proc. ACM Int. Conf. Knowl. Discov. Data Mining, 2002, pp. 475–480.

[8] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, "Privacy-preserving face recognition," in Proc. 9th Int. Symp. Privacy Enhancing Technol., 2009, pp. 235–253.

[9] R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving recordlinkage using Bloom filters," BMC Med. Inform. Decision Making,vol. 9, p. 41, 2009.

[10] R. Schnell, T. Bachteler, and J. Reiher, "A novel error-tolerantanonymous linking code," German Record Linkage Center,Working Paper Series No. WP-GRLC-2011-02, 2011.

[11] www.wikipedia.com/wiki