IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 5, Oct-Nov, 2013
ISSN: 2320 - 8791
www.ijreat.org

Efficient Techniques for Secure Record Linkage by Using Bloom Filters

P. Evangeline Jasmine[1], M.Nathiya[2] and R. Kavitha[3]

[1]PG Scholar, Department of CSE, VEL TECH UNIVERSITY, Avadi, Chennai-62.

[2]PG Scholar, Department of CSE, VEL TECH UNIVERSITY, Avadi, Chennai-62.

[3]Research Scholar, Department of CSE, VEL TECH UNIVERSITY, Avadi, Chennai-62.

Abstract

In recent years the need for consolidating the information contained in heterogeneous data sources has been widely documented. In order to achieve this goal, an organization must resolve several types of heterogeneity problems.Statistical record linkage techniques could be used for resolving this problems. However these techniques for onlinerecord linkage could pose a tremendous communication bottleneck in a distributed environment. In order to resolve this issue, we develop a matching tree, similar to a decision tree, and use it to proposetechniques that reduce the communication overhead significantly. When databases are maintained by disparate organizations, the disclosure of such information can breach the privacy of the corresponding individuals. Our objective is to adapt a Bloom Filter encoding technique to mitigate such attacks and we achieved the tradeoff between security and accuracy.

Keywords:*Entityheterogeneityproblem,Decision Tree,Bloom filter,data matching,record linkage,entity resolution,privacy,security.*

1 Introduction

The last few decades have witnessed a tremendous increase in the use of computerized databases forsupporting a variety of business decisions. The data needed to support these decisions are often scattered in heterogeneousdistributed databases. In such cases, it maybenecessary to link records in multiple databases so that onecan consolidate and use the data pertaining to the same realworldentity. If the databases use the same set of designstandards, this linking can easily be done using the primarykey (or other common candidate keys). However, since theseheterogeneous databases are usually designed and managedby different organizations (or different units within the same organization), there may be no common candidate key for linking the records. Although it may be possible to use common non key attributes (such as name, address, and dateof birth) for this purpose, the result obtained using theseattributes maynot always be accurate. This is because nonkeyattribute values may not match even when the recordsrepresent the same entity instance in reality. The above problem—where a real-world entity type isrepresented by different identifiers in two databases—isquite common in the real world and is called the entityheterogeneity problem or the common identifier problem .

The key question here is one of recordlinkage: given a record in a local database (often called theenquiry record), how do we find records from a remotedatabase that may match the enquiry record? Traditionalrecord linkage techniques, however, are designed to linkan enquiry record with a set of records in a local masterfile. Given the enquiry record and a record from the(local) master file, these techniques compare the commonnonkey attribute values of the two records to derive asimilarity measure—typically the probability of a matchor the likelihood ratio. If the similarity measure isabove a certain threshold, the two records are said tosatisfy the linkage rule. Record linkage techniques have been widely used inreal-world situations—such as health care[1],[2],[5] immigration and census where all the records are available locally.

However, when the matching records reside at a remotesite, existing techniques cannot be directly applied becausethey would involve transferring the entire remote relation,thereby incurring a huge communication overhead. As aresult, record linkage techniques do not have an efficientimplementation in an online.

In order to fully appreciate the overall difficulty,two important characteristics of the problem context mustbe understood:

*The databases exhibiting entity heterogeneity are distributed,and it is not possible to create and maintain acentral data repository or warehouse where pre-computedlinkage results can be stored.

*The participating sites allow controlled sharing of portionof their databases using standard database queries, but theydo not allow the processing of scripts, stored procedures, orother application programs from another organization.

When databases are maintained by disparate organizations, the disclosure of such information can breach the privacy of the corresponding individuals. Various private record linkage (PRL) methods have been developed to obscure such identifiers, but they vary widely in their abilityto balance competing goals of accuracy, efficiency and security. The tokenization and hashing of field values into Bloom filters (BF) enables greater linkage accuracy and efficiency than other PRL methods, but the encodings may be compromised through frequencybasedcryptanalysis. Our objective is to adapt a BF encoding technique to mitigate such attacks with minimal sacrifices in accuracyand efficiency. To accomplish these goals, we introduce a statistically-informed method to generate BF encodings that integrate bitsfrom multiple fields, the frequencies of which are provably associated with a minimum number of fields. Our method enables a userspecifiedtradeoff between security and accuracy.

1.1 Example: Crime Investigation

Consider the situation in a large metropolitan area consisting of about 40 municipal regions. Each municipality is equipped with (mostly incompatible) criminal data processingsystems and their respective data models. Although, the municipalities share a significant portion of the storedcriminal records among themselves, it has long beendecided that it is not practical to create a central datawarehouse that consolidates all the information.

Example, a police officer investigating a crime at the sitemakes a phone call to a backroom operator, who searchesthrough the different databases to determine if certainoffender types are known to be located in the call area ofinterest. The process is quite inefficient. First, it is oftendifficult for a police officer to relay the exact searchrequirements to the operator. Second, the police officer hasto rely on the operator’s expertise and intuition inmodifying the search criteria based on the results of aprevious query. Third, when the search criteria are satisfiedby several records in several databases, relaying all theinformation back to the police officer over the phone iscumbersome, error-prone, and time-consuming. Finally, ifall backroom operators are busy working on otherinvestigations, an officer may have to wait for a long timebefore an operator becomes available to provide thenecessary help.

In order to address this problem, a proposal is currentlyunder consideration whereby the field personnel (such asinvestigating officers, certain social workers, and forensicexperts) would be provided with handheld devices. Thebasic idea in this proposal is that a crime investigatorshould be able to quickly download relevant information (appropriate to the crime profile of the case at hand) onthese devices, instead of having to rely on a backroomoperator to do the necessary research.

Unfortunately, there are several challenges in implementing this proposal. First, since no centralized data warehouseexists, an investigating officer may have to send queries toseveral databases separately to download the relevantinformation. Second, the handheld devices do not haveenough storage capacity to download all the remotedatabases in a batch process and store them locally. Third,the connection speed on these machines (based on a wirelessnetworking infrastructure) is not very high, making itimpossible to download millions of records on a real-timebasis. Therefore, the practicality of the entire proposaldepends on finding a way to download only the relevantcriminal records to the handheld devices.

IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 5, Oct-Nov, 2013
ISSN: 2320 - 8791
www.ijreat.org

2. Proposed Model

In this section, we draw upon the research in the area ofsequential information acquisition [3], [4] to provide anefficient solution to the online, distributed record linkageproblem. The main benefit of the sequential approach isthat, unlike the traditional full-information case, not all theattributes of all the remote records are brought to the localsite; instead, attributes are brought one at a time. Afteracquiring an attribute, the matching probability is revisedbased on the realization of that attribute, and a decision ismade. For secure record linkage , we proposed RBF encoding will provide stronger resistanceagainst frequency analysis and therefore greatersecurity.

2.1 Sequential Record Linkage and Matching Tree

The sequential approach decides on the next "best"attribute to acquire, based upon the comparison results ofthe previously acquired attributes. The acquisition ofattributes can be expressed in the form of a matching treeas shown in fig 1.

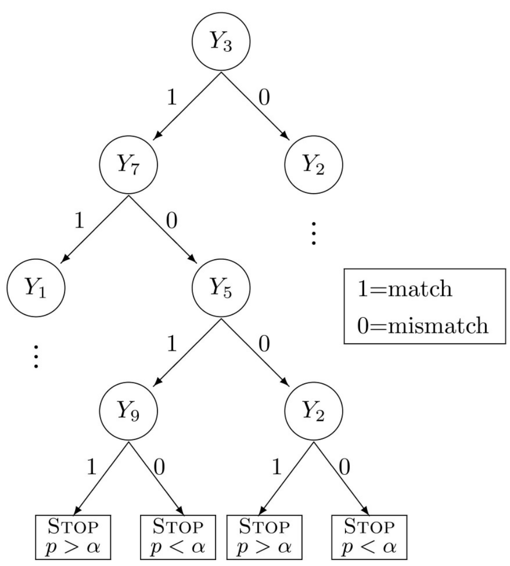

Fig. 1.A sample tree showing attribute acquisition order.

This tree can be used in the followingmanner: Starting at the root, we acquire attribute Y3 first. Ifthere is a match on this attribute, we acquire attribute Y7;otherwise, we acquire Y2. Similarly, after acquiring Y7, ifthere is a match, we acquire Y1, and so on, till a "STOP"node is reached. In the end, we would have a set ofprobability numbers for each remote record, based only ona subset of attributes that would have been acquired along apath of the tree. We now discuss how one can induce amatching tree similar to the one shown in Fig. 1.

There are two basic principles used in the induction of amatching tree: 1) input selection and 2) stopping. Before wedescribe these two principles, we would like to clarify animportant point. In inducing the tree, as well as in oursubsequent numerical analysis, we make the common assumption of conditional independence among Uks givenM; this reduces the overall computational burden. However,the idea presented here is more general, because, evenin situations where this assumption does not hold, thematching tree can still be constructed through recursivepartitioning of the training data, as is done in the traditional induction of a decision tree .

2.2 Tree Based Linkage Techniques

In this section, we develop efficient online record linkagetechniques based on the matching tree

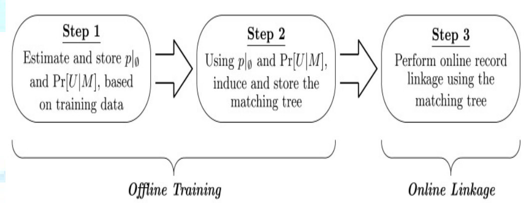

Fig. 2.The overall process of online tree-based linkage.

The overall linkage process is summarized in Fig. 2. The firsttwo stages in this process are performed offline, using thetraining data. Once the matching tree has been built, theonline linkage is done as the final step.We can now characterize the different techniques thatcan be employed in the last step. Recall that, given a localenquiry record, the ultimate goal of any linkage technique isto identify and fetch all the records from the remote site thathave a matching probability of one or more. In other words, one needs to partition the set of remote records into two subsets:1) relevant records that have a matching probability of one or more, and 2) irrelevant records that have a matching probability of less than one. Our aim is to develop techniquesthat would achieve this objective while keeping thecommunication overhead as low as possible. The partitioningitself can be done in one of two possible ways:**1) sequential, or 2) concurrent.**

In sequential partitioning, the set of remote records ispartitioned recursively, till we obtain the desired partitionof all the relevant records. This recursive partitioning can bedone in one of two ways: 1) by transferring the attributes ofthe remote records and comparing them locally, or 2) bysending a local attribute value, comparing it with the valuesof the remote records, and then transferring the identifiersof those remote records that match on the attribute value.we call the first one sequential attributeacquisition, and the second, sequential identifier acquisition. In the concurrent partitioning scheme, the tree is used toformulate a database query that selects the relevant remoterecords directly, in one single step.Hence, there is no needfor identifier transfer. Once the relevant records areidentified, all their attribute values are transferred.

In order to find the matching records, weimplement fuzzy matching for all the string-valued attributes. We first define a similarity measure between any two character strings τ_1 and τ_2, based on a character-by character comparison of these two strings:

$$\sigma(\tau_1, \tau_2) = \frac{1}{\mu} \sum_{i=1}^{\mu} I_{\tau_1[i]=\tau_2[i]},$$

Where

$I\tau_1[i]=\tau_2[i]$ is 1 only if the i^{th} characters of both the strings are the same, and it is zero otherwise; is the lengthof the shorter of the two strings.

3. Bloom Filter

A Bloom filter is a data structure for checking set membership efficiently. Bloom filters can also be used to determine whether two sets approximately match. If we want to compute the similarity between those strings without revealing the confidential data, we must use an encryption. Our protocol for privacy-preserving record linkage uses a Bloom filter for this task.

3.1 Contributions

The contributions of this work are:

1) **Enhanced security:** Our encoding method generates RBFs from FBF encodings via a data-driven bit selection procedure. This encoding utilizes a tunable security parameter with quantifiable resistance to frequency-based cryptanalysis attacks [13].

2) **Top Rank Preserving:** The resulting RBFs provide a transformation from the plaintext space to the cipher text space, such that the nearest neighbour to record is retained with a high likelihood. This paves the way for the application of PRL in the cipher text space in a manner that maintains a high degree of accuracy.

3) **Empirical Evaluation:** We perform an evaluation of the RBF strategy with several competing approaches using a dataset of personal identifiers derived from a real voter list. We use statistical hypothesis testing to demonstrate that the RBF strategy provides better top rank preservation than its competitors.

3.2 Record Level Bloom Filter

When databases are maintained by disparate organizations, the disclosure of such information can breach the privacy of the corresponding individuals. Various private record linkage (PRL) methods have been developed to obscure such identifiers, but they vary widely in their ability to balance competing goals of accuracy, efficiency and security.

The tokenization and hashing of field values into Bloom filters (BF)enables greater linkage accuracy and efficiency than other PRL methods, but the encodings may be compromised through frequencybasedcryptanalysis. RBF encoding will provide stronger resistanceagainst frequency analysis and therefore greatersecurity.

4 Conclusions

In this paper, we develop efficient techniques to facilitaterecord linkage decisions in a distributed, online setting.Record linkage is an important issue in heterogeneousdatabase systems where the records representing the samereal-world entity type are identified using different identifiers
in different databases.

To accomplish the security issues in online record linkage we have adopted bloom filters to enhance the accuracy and efficiency for consolidating the heterogenous data sources.

References

[1] J.A. Baldwin, "Linked Record Health Data Systems," TheStatistician, vol. 21, no. 4, pp. 325-338, 1972.

[2] M.J. Goldacre, J.D. Abisgold, D.G.R. Yeates, and V. Seagroatt,"Risk of Multiple Sclerosis after Head Injury: Record LinkageStudy," J. Neurology, Neurosurgery, and Psychiatry, vol. 77, no. 3,pp. 351-353, 2006.

[3] J. Moore and A. Whinston, "A Model of Decision Making withSequential Information Acquisition—Part I," Decision SupportSystems, vol. 2, no. 4, pp. 285-307, 1986.

[4] J. Moore and A. Whinston, "A Model of Decision Making withSequential Information Acquisition—Part II," Decision SupportSystems, vol. 3, no. 1, pp. 47-72, 1987.

[5] A.M. Ward, N. de Klerk, D. Pritchard, M. Firth, and C.D. Holman,"Correlations of Siblings' and Mothers' Utilization of Primary andHospital Health Care: A Record Linkage Study in Western
Australia," Social Science and Medicine, vol. 62, no. 6, pp. 1341-1348, 2005.

[6] 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.