Efficient Techniques for Secure Record Linkage by Using Bloom Filters

P. Evangeline Jasmine¹, M. Nathiya² and R. Kavitha³

¹PG Scholar, Department of CSE, VEL TECH UNIVERSITY, Avadi, Chennai-62.
²PG Scholar, Department of CSE, VEL TECH UNIVERSITY, Avadi, Chennai-62.
³Research 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 propposeotechniques 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: Entity heterogeneity problem, 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 for supporting a variety of business decisions. The data needed to support these decisions are often scattered in heterogeneous distributed databases. In such cases, it may be necessary to link records in multiple databases so that one can consolidate and use the data pertaining to the same real-world entity. If the databases use the same set of design standards, this linking can easily be done using the primary key (or other common candidate keys). However, since these heterogeneous databases are usually designed and managed by 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 date of birth) for this purpose, the result obtained using these attributes may not always be accurate. This is because nonkeyattribute values may not match even when the records represent the same entity instance in reality. The above problem—where a real-world entity type is represented by different identifiers in two databases—is quite common in the real world and is called the entity heterogeneity problem or the common identifier problem.

The key question here is one of record linkage: given a record in a local database (often called the enquiry record), how do we find records from a remote database that may match the enquiry record? Traditional record linkage techniques, however, are designed to link an enquiry record with a record from the local master file. Given the enquiry record and a record from the (local) master file, these techniques compare the common non-key attribute values of the two records to derive a similarity measure—typically the probability of a match or the likelihood ratio. If the similarity score is above a certain threshold, the two records are said to satisfy the linkage rule. Record linkage techniques have been widely used in real-world situations—such as health care[1],[2],[5] and census where all the records are available locally.

However, when the matching records reside at a remote site, existing techniques cannot be directly
applied because they would involve transferring the entire remote relation, thereby incurring a huge communication overhead. As a result, record linkage techniques do not have an efficient implementation in an online.

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

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

* The participating sites allow controlled sharing of portions of their databases using standard database queries, but they do not allow the processing of scripts, stored procedures, or other 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 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 frequency-based cryptanalysis. Our objective is to adapt a BF encoding technique to mitigate such attacks with minimal sacrifices in accuracy and efficiency. To accomplish these goals, we introduce a statistically-informed method to generate BF encodings that integrate bits from multiple fields, the frequencies of which are provably associated with a minimum number of fields. Our method enables a user-specified tradeoff 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 processing systems and their respective data models. Although, the municipalities share a significant portion of the stored criminal records among themselves, it has long been decided that it is not practical to create a central data warehouse that consolidates all the information.

Example, a police officer investigating a crime at the site makes a phone call to a backroom operator, who searches through the different databases to determine if certain offender types are known to be located in the call area of interest. The process is quite inefficient. First, it is often difficult for a police officer to relay the exact search requirements to the operator. Second, the police officer has to rely on the operator’s expertise and intuition in modifying the search criteria based on the results of a previous query. Third, when the search criteria are satisfied by several records in several databases, relaying all the information back to the police officer over the phone is cumbersome, error-prone, and time-consuming. Finally, if all backroom operators are busy working on other investigations, an officer may have to wait for a long time before an operator becomes available to provide the necessary help.

In order to address this problem, a proposal is currently under consideration whereby the field personnel (such as investigating officers, certain social workers, and forensic experts) would be provided with handheld devices. The basic idea in this proposal is that a crime investigator should be able to quickly download relevant information (appropriate to the crime profile of the case at hand) on these devices, instead of having to rely on a backroom operator to do the necessary research.

Unfortunately, there are several challenges in implementing this proposal. First, since no centralized data warehouse exists, an investigating officer may have to send queries to several databases separately to download the relevant information. Second, the handheld devices do not have enough storage capacity to download all the remote databases in a batch process and store them locally. Third, the connection speed on these machines (based on a wireless networking infrastructure) is not very high, making it impossible to download millions of records on a real-time basis. Therefore, the practicality of the entire proposal depends on finding a way to download only the relevant criminal records to the handheld devices.
2. Proposed Model

In this section, we draw upon the research in the area of sequential information acquisition [3], [4] to provide an efficient solution to the online, distributed record linkage problem. The main benefit of the sequential approach is that, unlike the traditional full-information case, not all the attributes of all the remote records are brought to the local site; instead, attributes are brought one at a time. After acquiring an attribute, the matching probability is revised based on the realization of that attribute, and a decision is made. For secure record linkage, we proposed RBF encoding will provide stronger resistance against frequency analysis and therefore greater security.

2.1 Sequential Record Linkage and Matching Tree

The sequential approach decides on the next “best” attribute to acquire, based upon the comparison results of the previously acquired attributes. The acquisition of attributes can be expressed in the form of a matching tree as shown in Fig. 1.

This tree can be used in the following manner: Starting at the root, we acquire attribute Y3 first. If there is a match on this attribute, we acquire attribute Y7; otherwise, we acquire Y2. Similarly, after acquiring Y7, if there is a match, we acquire Y1, and so on, till a “STOP” node is reached. In the end, we would have a set of probability numbers for each remote record, based only on a subset of attributes that would have been acquired along a path of the tree. We now discuss how one can induce a matching tree similar to the one shown in Fig. 1.

There are two basic principles used in the induction of a matching tree: 1) input selection and 2) stopping. Before we describe these two principles, we would like to clarify an important point. In inducing the tree, as well as in our subsequent numerical analysis, we make the common assumption of conditional independence among Uks given M; this reduces the overall computational burden. However, the idea presented here is more general, because, even in situations where this assumption does not hold, the matching tree can still be constructed through recursive partitioning 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 linkage techniques based on the matching tree.

The overall linkage process is summarized in Fig. 2. The first two stages in this process are performed offline, using the training data. Once the matching tree has been built, the online linkage is done as the final step. We can now characterize the different techniques that can be employed in the last step. Recall that, given a local enquiry record, the ultimate goal of any linkage technique is to identify and fetch all the records from the remote site that have 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 techniques that would achieve this objective while keeping the communication overhead as low as possible. The
partitioning itself can be done in one of two possible ways: 1) sequential, or 2) concurrent.

In sequential partitioning, the set of remote records is partitioned recursively, till we obtain the desired partition of all the relevant records. This recursive partitioning can be done in one of two ways: 1) by transferring the attributes of the remote records and comparing them locally, or 2) by sending a local attribute value, comparing it with the values of the remote records, and then transferring the identifiers of those remote records that match on the attribute value. We call the first one sequential attribute acquisition, and the second, sequential identifier acquisition. In the concurrent partitioning scheme, the tree is used to formulate a database query that selects the relevant remote records directly, in one single step. Hence, there is no need for identifier transfer. Once the relevant records are identified, all their attribute values are transferred.

In order to find the matching records, we implement fuzzy matching for all the string-valued attributes. We first define a similarity measure between any two character strings \( \tau_1 \) and \( \tau_2 \), 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; \( \mu \) is the length of 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 ciphertext space, such that the nearest neighbor to record is retained with a high likelihood. This paves the way for the application of PRL in the ciphertext 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 frequency-based cryptanalysis. RBF encoding will provide stronger resistance against frequency analysis and therefore greater security.

4 Conclusions

In this paper, we develop efficient techniques to facilitate record linkage decisions in a distributed, online setting. Record linkage is an important issue in heterogeneous database systems where the records representing the same real-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 heterogeneous data sources.
References


