Privacy Preserving Blocking and Meta-Blocking

Alexandros Karakasidis[1(✉)], Georgia Koloniari[2], and Vassilios S. Verykios[1]

[1] School of Science & Technology, Hellenic Open University, Patras, Greece
{a.karakasidis,verykios}@eap.gr
[2] Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece
gkoloniari@uom.gr

Abstract. Record linkage refers to integrating data from heterogeneous sources to identify information regarding the same entity and provides the basis for sophisticated data mining. When privacy restrictions apply, the data sources may only have access to the merged records of the linkage process, comprising the problem of privacy preserving record linkage. As data are often dirty, and there are no common unique identifiers, the linkage process requires approximate matching and it renders to a very resource demanding task especially for large volumes of data. To speed up the linkage process, privacy preserving blocking and meta-blocking techniques are deployed. Such techniques derive groups of records that are more likely to match with each other. In this nectar paper, we summarize our contributions to privacy preserving blocking and meta-blocking.

Keywords: Privacy · Record linkage · Blocking · Meta-blocking

1 Introduction

Considering the data explosion we experience the last decade, we seek ways to boost the results of data mining. As related data are highly scattered, i.e., in different organizations databases, on the web, etc., integrating large volumes of data comprises an indispensable first step towards mining more useful information that could not be discovered if we consider each separate database in isolation.

This process of identifying and linking information across multiple databases, that refers to the same real world entity, is known as the problem of *record linkage*. When privacy concerns arise, the record linkage problem is augmented to its privacy preserving version, where the participants should not gain any additional information regarding each other's data, apart from the linkage results.

For instance, let us consider a medical researcher who wishes to perform a study on the interactions between certain prescribed medicine over the last decade, using data from hospitals and clinics from all over Europe. This comprises a data mining problem, where the additional requirement of privacy is posed due to the sensitive nature of the data. These data are not all stored in a single database which may be mined, but originate from multiple health care

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 232–236, 2015.
DOI: 10.1007/978-3-319-23461-8_20

Fig. 1. Privacy preserving record linkage workflow.

units from different countries each of them using its own database. Consequently, these data should be merged, using a private record linkage protocol.

The lack of global unique identifiers deems necessary the use of common attributes, in most cases textual, for identifying the matching records. As attribute values are usually the result of user input, they are most often dirty, requiring methods for approximate matching. Taking into account the large volumes of available data, we confront a very resource demanding task. To deal with this, *privacy preserving matching* (PPM) is often preceded by a *privacy preserving blocking* (PPB) phase. PPB speeds up matching by organizing candidate records that are more likely to match into blocks, based on the values of their attributes. The attributes selected for PPB and PPM are respectively called blocking and matching attributes. Lately, privacy preserving meta-blocking was introduced which, applied after blocking, aims at reorganizing the way records within a block should be matched, so as to further improve performance. The phases of the overall linkage process are depicted in Fig.1.

Blocking imposes an additional filtering step to the matching process, thus increasing its precision. On the other hand, blocking may eliminate matching record pairs, thus decreasing recall. Therefore, some blocking techniques compromise result quality [9], while others rely on efficiency-privacy tradeoffs failing to significantly improve performance for large scale data without sacrificing their privacy [8]. Finally, there are approaches that, though efficient, are limited to specific data types, either numerical or nominal [2]. In this nectar paper, we present our contribution on privacy preserving blocking and meta-blocking methods. Our aim is to boost performance while maintaining high levels of matching quality without compromising privacy.

2 Privacy Preserving Blocking and Meta-Blocking

We first present three privacy preserving blocking techniques and then, the only work up to now on privacy preserving meta-blocking. The first blocking technique is designed for textual data, while all others may be adopted for both textual and numerical data using appropriate distance and similarity measures.

Phonetic Code Based Private Blocking. A phonetic code is a hash produced by a phonetic algorithm for matching words based on their pronunciation. The main feature of phonetic algorithms is their fault tolerance against typographical errors. In [3], we present a two-party phonetic based privacy preserving

blocking protocol. The two parties (data sources) agree on the use of a set of phonetic algorithms and blocking attributes. Each party then encodes the blocking fields with each of the algorithms. To increase the entropy of each dataset and consequently reduce the ability of predicting its values, fake phonetic codes are injected. Next, all phonetic codes are encrypted using a secure hash function and records are grouped into blocks according to their hashes. Each record is assigned to multiple blocks according to each of the blocking attributes encoded by each of the phonetic algorithms.

High matching quality is assured by using multiple phonetic codes per blocking attribute, thus overcoming through redundancy certain weak points of phonetic algorithms, such as in Soundex [7], where an error in the first letter produces a different code. Privacy is assured as phonetic algorithms are one-way functions which apply information suppression, and improved with the use of fake codes and encryption. With respect to efficiency, phonetic codes are very fast to compute, and moreover, matching on identical phonetic hashes enables us to deploy indexes to further speed up matching, achieving up to 61.4% speedup with respect to plain matching and recall at 0.67 [3].

Reference Table Based Private Blocking. Reference tables are publicly available datasets used to provide privacy by avoiding direct comparison between the two databases, using instead reference values as a comparison basis. Our contribution comprises of two methods, that employ a third party. The two sources individually cluster the same reference values. Each record is classified to a cluster (class) based on some distance or similarity measure, associating one of the reference values to its blocking attribute. Records classified at the same class comprise a block. The two sources send the classified record ids to the third party who merges blocks belonging to the same class. Final blocks are returned to the sources only when they contain record ids from both sources. Matching may then be performed either at the sources or at the third party.

Reference table based k-anonymous private blocking [5] is the first work using this concept for privacy preserving blocking. Nearest Neighbor clustering is used to form clusters of at least k-elements, thus ensuring k-anonymity as each record is assigned to a class based on its similarity with one of at least k-elements. However, while ensuring privacy and result quality, the method incurs high complexity as each blocking attribute should be looked up against all reference values. Experiments show that linkage with our method requires half the execution time of plain matching with recall up to 0.78 [5].

Multidimensional private blocking [4] improves the performance of [5], by using k-Medoids for clustering. Each blocking attribute is checked only against cluster medoids, thus reducing the method's complexity. Moreover, the use of an edit distance negates the need for a reference table to contain values similar to the ones contained in the datasets. For numerical fields, bins are created based on numerical reference values. This work introduces the concept of multidimensional blocking. In blocking, when more than one blocking fields are used, the same procedure has to be repeated and a record may fall within numerous blocks.

With multidimensional blocking, the two sources locate the class of each record for each of the blocking attributes used. Then, they calculate the intersection of classes each record belongs to. As such, a record is associated with less blocks resulting in reduced matching operations. Execution time drops further to 11% of plain matching, while recall is 0.73 [4].

Privacy Preserving Meta-Blocking. *Sorted neighborhood on encrypted fields* (SNEF) [6], based on [1], is to the best of our knowledge, the only privacy preserving meta-blocking method. Multidimensional private blocking is extended by associating each record within each block with a score derived by an objective function that uses the edit distance between each blocking attribute and the cluster medoid of its class. After the third party merges the blocks, the party who performs the privacy preserving matching sorts the records within each block based on their scores. A sliding window of size w slides over the records, and each record is checked against the next w records in the block, rendering the matching complexity within each block from quadratic to linear. SNEF does not compromise privacy since each record is associated with a single number which cannot be factorized due to the properties of the objective function. There is a tradeoff between matching quality and time efficiency, depending on window size which, nevertheless, remains linear. As experiments show, SNEF further improves multidimensional blocking's time by 20% with a recall around 0.70 [6].

3 Conclusion and Future Work

We presented our work on privacy preserving blocking techniques, which are efficient while assuring privacy and result quality. Next, we plan to accelerate our blocking methods by using random samples instead of clustering.

References

1. Hernández, M.A., Stolfo, S.J.: Real-world data is dirty: Data cleansing and the merge/purge problem. Data Min. Knowl. Discov. **2**(1), 9–37 (1998)
2. Inan, A., Kantarcioglu, M., Ghinita, G., Bertino, E.: Private record matching using differential privacy. In: ACM EDBT (2010)
3. Karakasidis, A., Verykios, V.S.: Secure blocking + secure matching = secure record linkage. JCSE **5**(3), 223–235 (2011)
4. Karakasidis, A., Verykios, V.S.: A highly efficient and secure multidimensional blocking approach for private record linkage. In: IEEE ICTAI (2012)
5. Karakasidis, A., Verykios, V.S.: Reference table based k-anonymous private blocking. In: ACM SAC (2012)

6. Karakasidis, A., Verykios, V.S.: A sorted neighborhood approach to multidimensional privacy preserving blocking. In: IEEE ICDMW (2012)
7. Odell, M., Russell, R.: The Soundex coding system. US Patents 1261167 (1918)
8. Vatsalan, D., Christen, P., Verykios, V.S.: Efficient two-party private blocking based on sorted nearest neighborhood clustering. In: ACM CIKM (2013)
9. Yakout, M., Atallah, M.J., Elmagarmid, A.K.: Efficient private record linkage. In: IEEE ICDE (2009)