Extended abstract - Joint Statistical Meetings (JSM), Montreal, August 2013

Overview and taxonomy of techniques for privacy-preserving record linkage

Assoc Prof Peter Christen
Research School of Computer Science, The Australian National University,
Canberra, ACT 0200, Australia
peter.christen@anu.edu.au

Abstract

Record linkage is the process of identifying which records in two or more databases correspond to the same real-world entity. Three major challenges of this process are (1) achieving high linkage quality, (2) scalability to linking very large databases, and (3) protecting the privacy and confidentiality of personal identifying data that are used in the linkage process.

This presentation provides an overview of the various techniques that have been developed to facilitate the linking of data across organisations in such ways that no private or confidential information is being revealed. We then characterise such privacy-preserving record linkage techniques along fifteen dimensions. This provides us with a taxonomy that allows us to highlight shortcomings of current techniques and discuss future research directions.

Background

Over the past decade there has been an increased interest in research domains that deal with the management, processing, and analysis of increasingly large data collections. A particular direction of research has focused on algorithms and techniques that allow the sharing and analysis of large data collections in such ways that the privacy and confidentiality of the shared and analysed data, as well as the knowledge gained from them, is maintained [2, 5].

Sharing, integrating, and linking disparate databases is an important aspect in the initial stages of many data analysis projects. This process is commonly known as *record linkage*, *data matching*, *entity resolution* or *duplicate detection* [1, 3]. The aim of this process is to identify which records in these databases refer to the same real-world entities. This allows improvements of data quality (for example by removing duplicate records that refer to the same entity) and facilitates data enrichment (by combining complementary information stored in different databases).

When record linkage is conducted between different organisations, then the privacy and confidentiality of the data required for the linking become a major concern. Often, it is not permissible to exchange identifying information (such as people's personal details) across different organisations, either because of legal regulations or because the data are of confidential nature.

Record linkage poses several major challenges [3]. The lack of common unique entity identifiers means that the identification of common entities needs to be based on the available information that is shared among records, such as the names, addresses, and dates of birth of individuals. The challenge here is poor data quality, which is tackled by the development of advanced similarity functions that try to distinguish different forms of the same information from actual different information. The second major challenge is the complexity of the linkage process when each record from one database is compared with all records from another database. This challenge is addressed by sophisticated searching and indexing techniques that reduce the large space of record comparisons through blocking, clustering or sorting of the databases. Research is conducted to address both these challenges to record linkage [1, 3].

Privacy-preserving record linkage

The third major challenge of record linkage, privacy and confidentiality, is addressed by the research area of *privacy-preserving record linkage* (PPRL). PPRL aims to develop algorithms and techniques that can identify records that refer to the same real-world entities from two or more databases owned by different organisations, such that besides the matched records no private or confidential information is revealed. A challenging aspect of PPRL is that it needs to deal with the previously described two challenges as well.

This presentation starts with providing the necessary background on record linkage, including a short history, its applications, and challenges. It then moves on to describe the record linkage and PPRL process, and illustrates the importance of PPRL using two example scenarios.

The second part of the presentation consists of an overview of our recently developed taxonomy for PPRL [4], where we characterised PPRL techniques along fifteen dimensions grouped into the following five topics:

- Privacy aspects: The number of parties involved in a PPRL protocol, the adversary model assumed (malicious or honest-but-curious), and which of the many available privacy technologies are employed.
- Linkage techniques: The approaches used for indexing (blocking), matching (comparisons), and classification.
- Theoretical analysis of a PPRL protocol: Scalability, linkage quality, and privacy vulnerabilities (with regard to different types of attack).
- (Experimental) evaluation: Scalability, linkage quality, and privacy.
- Practical aspects: Implementation, data sets (used for experiments), application area.

Each of these five topics is briefly presented and discussed. The interested reader is referred to [4] for more details. The presentation concludes with a summary of the current state of PPRL and a description of open research questions and directions. The main research directions we have identified are (1) improved classification techniques for PPRL; (2) assessing linkage quality and completeness for PPRL; (3) a framework for PPRL (to facilitate comparative experimental evaluations); and (4) enabling efficient PPRL on multiple large databases.

References

[1] Christen P: *Data matching – Concepts and techniques for record linkage, entity resolution, and duplicate detection*, Springer *Data-centric systems and applications*, 2012.

[2] Clifton C, Kantarcioğlu M, Doan A, Schadow G, Vaidya J, Elmagarmid A and Suciu D: *Privacy-preserving data integration and sharing.* ACM SIGMOD workshop on Research Issues in Data Mining and Knowledge Discovery, Paris, pp. 19–26, 2004.

[3] Elmagarmid AK, Ipeirotis PG and Verykios VS: *Duplicate record detection: A survey.* IEEE TKDE 19(1), pp. 1–16, 2007.

[4] Vatsalan D, Christen P and Verykios VS: *A taxonomy of privacy-preserving record linkage techniques.* Elsevier Information Systems, (38)6, pp. 946-969, 2013.

[5] Verykios VS, Bertino E, Fovino IN, Provenza LP, Saygin Y and Theodoridis Y: *State-of-the-art in privacy preserving data mining.* SIGMOD Rec, (33)1, pp. 50–57, 2004.