Privacy Preserving Record Linkage: Techniques and Experimental Evaluations

Dr. Li Xiong
Emory University
Li.Xiong@emory.edu

Courtney A. Beach
Albany State University
cbeach@students.asurams.edu

From this internship, I have learned that record linkage is the matching of two records with the same or similar attributes such as identification number or address. Record Linkage is the computation of the associations among records of multiple databases. Private Record Linkage is the problem of carrying out the linkage computation without full data exchange. Privacy preserving record linkage allows two parties to identify records that are closely related to each other using a distances formula such that no additional information, other than the results are disclosed to each party. Another reason it can be useful for consolidating patient's records between hospitals. The goal of my research was to learn about both record linkage, privacy preserving record linkage, and to use taxonomy for classifying the different techniques, give an overview of the representative techniques, and present comparative analysis and experimental findings regarding their security, accuracy and performance.

There were about 5 different papers I had to read to gain knowledge about privacy preserving record linkage. I wrote summaries on 3 of the 5 papers about the various ways to link records. All of them offered different approaches to solving the problem of preserving privacy while linking records. In the BMC Medical Informatics and Decision Making Some methods for blindfolded record linkage paper, the issue was how to perform minimal-knowledge similarity comparisons between strings, using keyed hash combinations of n-grams. Blindfolded Record Linkage is a secure one-way hash transformation to carry out follow-up epidemiological studies without any party having to reveal identifying information about any of the subjects. The method proposed to solve the issue permits the calculation of general similarity measure. This method can be combined with public key cryptography and automatic estimation of linkage model parameters to create a system to blindfolded record linkage.

On another note, Privacy Preserving Schema and Data Matching focuses on record matching performed with the aim of identifying common information shared by two data sources. The goal of the paper is to allow each source or party the ability to hide their records not to be shared with the other source, the detail of the attributes in its schema and several other features that the source may want to keep private. The technique proposed for privacy preserving record linkage does not rely on the usage of complex cryptographic processes but assures

privacy by application of a method widely used for similarity based searching of complex objects. The main idea of our work is to embed records to be matched in an Euclidean space and to perform the comparison. The protocol is proposed for record matching that preserves privacy both at the data level and at the schema level. It is a privacy-preserving protocol to perform record matching across two data sources. The protocol allows each source to hide the records not to be shared with the other source, the detail of the attributes in its schema, and several other features that the source may want to keep private. This particular protocol performs privacy-preserving approximate matching.

The main issue of Efficient Private Record Linkage is how to carry out record linkage while maintaining privacy? The proposed solution is to build a matching function that identifies the matched records and operates in a privacy preserving manner. The techniques proposed for the solution in this work are supposed to improve on previous works in that (1) they make no use of a third party and (2) they achieve much better performance than that of previous schemes in terms of execution time and quality of output. To solve the problem of privacy record linkage, the researchers propose use of cryptographic methods; apply perturbation methods on private information to obscure individual identity; and non-cryptographic techniques using an outside third party. The cryptographic technique guarantees accurate results with high privacy but it is not practical to be used for large databases or approximating matches. Perturbation methods are cost efficient but lack the required accuracy for linking records. Non-cryptographic techniques do not require either party to be aware of each other due to a third party linking the information. The proposed protocol for Efficient Private Record Linkage has two phases, with the first phase being the main goal. Phase 1 is to produce candidate pairs of records for matching, by carrying out a very fast but not accurate matching between pairs of records. Phase 2 uses a practical privacy-preserving protocol for completing the task of computing the Euclidean distance between each candidate pair. Both parties participate in its computations without revealing the original vector representations of their respective records.

Overall, my research experience at Emory University with Dr. Li Xiong was a learning experience. I learned about a new topic in computer that pertains to my two favorite subjects, database management and computer security. If time would permit I would have like to continue the research or a least learn more about the subject of privacy preserving record linkage. I did learn about other things that will help me in my academic career such as how to navigate around the UNIX operating system, how to use Eclipse for java programming and how to effectively and efficiently read a research paper. I also learned a lot about Georgia highways and how to balance family life and business. Atlanta was a new experience for me because I'm from a much smaller city. The best part of my entire summer experiences is that I remained optimistic about every situation that was brought before me.

References:

1. T. Churches and P. Christen. Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making, 4(9), 2004.

2. M. Scannapieco, I. Figotin, E. Bertino, and A.K. Elmagarmid. Privacy preserving schema and data matching. In S:IGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pages 653-664, New York, NY, USA, 2007, ACM.

3. M.Yakout, M.J. Atallah, and A.K. Elmagarmid, Efficient private record linkage. In ICDE, pages 1283-1286, 2009.