Privacy-Preserving Schema Matching Using Mutual Information*

Isabel F. Cruz[1], Roberto Tamassia[2], and Danfeng Yao[2]

[1] Department of Computer Science
University of Illinois at Chicago
ifc@cs.uic.edu
[2] Department of Computer Science
Brown University
{rt,dyao}@cs.brown.edu

The problem of *schema or ontology matching* is to define *mappings* among schema or ontology elements. Such mappings are typically defined between two schemas or two ontologies at a time. Ideally, using the defined mappings, one would be able to issue a single query that will be rewritten automatically to all the databases, instead of manually writing a query to each database. In a centrally mediated architecture a query is written in terms of a global schema or ontology that integrates all the database schemas or ontologies, while in a peer-to-peer architecture a query is written in terms of the schema or of the ontology of any of the peer databases.

Automatic schema matching approaches can use only the schema, only the instances, or a combination of both. Mappings can take into account not only concept properties (e.g., string similarity), but also constraints (e.g., relationship cardinality) and schema structure (e.g., graph similarity) [9].

Security and privacy issues arise in the context of data integration. For example, previous work looks into secure access to mediated data [2, 4]. Other work has defined the concept of *minimal necessary information sharing* that applies to querying: in computing the answer to a query, only the query result should be revealed [1]. Most matching approaches rely on the fact that both schemas or ontologies are completely visible by both parties. Clearly, this approach disregards security and privacy considerations. Even within the same organization, different users have access to different database views. It is, therefore, only natural to create automatic mechanisms by which mappings can be established between a pair of schemas or ontologies, without each party needing to reveal their whole metadata.

Clifton *et al.* discuss issues and identify research directions in privacy-preserving data integration, including those that arise in schema matching [3]. More recently, Mitra *et al.* look at the specific issue of privacy-preserving ontology matching [7, 8]. In their approach, terms in the ontologies and in the matching rules (which define the mappings) are encrypted, so that the mediator does not see the actual terms. However, during the ontology matching process, which is semi-automatic, a human expert has access to both ontologies in cleartext (using a session key).

We propose an automatic privacy-preserving schema matching protocol. The result of this protocol is the set of mappings between attributes in the schemas of the two inter-

* This work was supported in part by the National Science Foundation under ITR awards IIS–0326284, IIS–0324846, and IIS–0513553.

vening parties. Most importantly, from a privacy-preserving viewpoint, we do not use a third-party mediator and only those schema attributes that are matched are revealed by a party to the other party.

Our approach to privacy-preserving schema matching is based on the instance-based schema matching approach by Kang and Naughton [6], which considers the dependencies among data instances, as measured by the mutual information among every pair of attributes in each schema. For each schema, these dependencies are represented as a weighted graph and matching between the two schemas relies on matching the corresponding graphs. The mutual information between two attributes is a measure of the amount of information that each attribute contains about the other attribute. Mutual information can be computed using the entropies of the individual attributes and the conditional entropies. We consider three types of mappings: one-to-one, onto, and partial.

We develop an efficient privacy-preserving schema matching protocol using mutual information of pair-wise attributes. The protocol is executed by two entities, each having a private schema. The output of the protocol is a set of mappings between the matching attributes of the two schemas. We prove that our privacy-preserving schema matching protocol is secure against malicious adversaries for all mapping types. One of the building blocks of our protocol is the privacy-preserving set intersection scheme by Freedman, Nissim, and Pinkas [5]. We show that in the case where all the attribute entropies in one of the schemas are different from one another, the protocol executes a linear number of privacy-preserving set intersections.

References

1. R. Agrawal, A. V. Evfimievski, and R. Srikant. Information sharing across private databases. In *Proc. ACM SIGMOD*, pages 86–97, 2003.
2. K. S. Candan, S. Jajodia, and V. S. Subrahmanian. Secure mediated databases. In *Proc. IEEE Int. Conf. on Data Engineering*, pages 28–37, 1996.
3. C. Clifton, M. Kantarcioglu, A. Doan, G. Schadow, J. Vaidya, A. K. Elmagarmid, and D. Suciu. Privacy-preserving data integration and sharing. In *Proc. ACM Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 19–26, 2004.
4. S. Dawson, S. Qian, and P. Samarati. Providing security and interoperation of heterogeneous systems. *Journal of Distributed and Parallel Databases*, 8(1):119–145, 2000.
5. M. Freedman, K. Nissim, and B. Pinkas. Efficient private matching and set intersection. In *Advances in Cryptology—Eurocrypt*, vol. 3027 of *LNCS*, pages 1–19. Springer, 2004.
6. J. Kang and J. F. Naughton. On schema matching with opaque column names and data values. In *Proc. ACM SIGMOD*, pages 205–216. 2003.
7. P. Mitra, P. Liu, and C.-C. Pan. Privacy-preserving ontology mapping. In *Proc. Int. Workshop on Contexts and Ontologies: Theory, Practice and Applications*, 2005.
8. P. Mitra, C.-C. Pan, P. Liu, and V. Atluri. Privacy-preserving semantic interoperation and access control of heterogeneous databases. In *Proc. ACM Conf. on Computer and Communications Security*, pages 66–77, 2006.
9. E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.