Privacy-preserving Ontology Matching

Prasenjit Mitra, Peng Liu, Chi-Chun Pan

The Pennsylvania State University,
University Park, PA 16802
{pmitra, pliu, cpan}@ist.psu.edu

Abstract

Increasingly, there is a recognized need for secure information sharing. In order to implement information sharing between diverse organizations, we need privacy-preserving interoperation systems. In this work, we describe two frameworks for privacy-preserving interoperation systems. Ontology matching is an indispensable component of interoperation systems. To implement privacy-preserving interoperation systems, we need privacy-preserving ontology matching algorithms. In this paper, we outline frameworks for privacy-preserving ontology matching and discuss the privacy implications of the frameworks.

Introduction

Though researchers have built tools that enable organizations to share information, largely, most of these tools have not taken into the account the necessity of maintaining the privacy and confidentiality of the data and the metadata of the organizations that want to share information.

Consider the (hypothetical, but seemingly probable) scenario where the U.S. and U.K. military want to share information. They want to share data only about the mission at hand while preserving the privacy of their systems. That is, they want to share information without exposing to each other any significant details about the schema and other metadata about their systems. To the best of our knowledge, the current state-of-the-art systems do not allow privacy-preserving information sharing without sharing that is required in such a scenario.

Not only does the need for secure information sharing arise among organizations that want to share information among each other, but the need also arises for intra-organization information sharing. Large organizations, like large corporations or even the U.S. Department of Homeland Security, have a number of departments with varying levels of autonomy. That is, even within the same organization, different departments use information systems that were autonomously constructed. For example, in a large software development firm, the data center may be located at a different geographical location than the software development department, and due to their different needs, the two departments maintain different systems. The challenge of secure information sharing is prevalent even in these scenarios.

Not only must an organization preserve the privacy of its data, but it must also preserve the privacy of sensitive metadata (or meta-information). Metadata describes how data are organized in the organization (e.g., data schema), how accesses are controlled in the organization (e.g., the internal access control policy and role hierarchies), and the semantics of the data used in the organization (e.g., ontology).

Organizations seeking to interoperate are increasingly using metadata like ontologies to capture the semantics of terms used in the information sources maintained by the organizations. Traditionally, it has been assumed that these ontologies will be published by the organization. Published ontologies from different organizations are matched and matching rules generated. Queries to information sources are rewritten using these matching rules so that the vocabulary used in the query is the same as that used by the information source.

Unlike in the traditional scenario, some organizations do not want to publish their metadata or even share the metadata with external users. Yet, they want to enable interoperation. In this scenario, the privacy of the metadata, e.g., the ontologies of information sources or the schema of databases, must be preserved. That is, any user outside the host organization should not have access to the ontologies in cleartext. This is because in a mediated architecture, if the mediator is malicious or if an intruder breaks in to the mediator, substantial loss of information and privacy occurs.

In this paper, we present two frameworks for privacy-preserving interoperation and especially highlight their privacy-preserving ontology-matching components. These frameworks achieve ontology matching with minimal "privacy leak" of the ontologies being matched. The interoperation system does not assume a trusted mediator. Ideally, the organizations want the mediator to gain minimal information about the data and the metadata stored in its information sources. In our system, the mediator operates over encrypted queries, encrypted ontologies and encrypted data.

In the first ontology-matching framework, we show how totally automated ontology mapping can be achieved. In this framework, the queries and ontologies are encrypted using a symmetric private key shared between the organizations interoperating. In the second framework, we show how semi-automatic ontology mapping can be achieved. In this framework, the ontologies of each organization are encrypted using their own private keys and thus even interoperating organizations do not share their ontologies. To the best of our knowledge, there exists no existing work on privacy-preserving ontology matching.

The difficulty in preserving the privacy of the ontologies is that totally automated ontology matching does not work very well in practice (despite individual claims in research settings). Even if automated methods achieve about 70-80% accuracy, the matching rules missed by the automated matchers must be generated manually. Now, in order for a human expert to match the ontologies to generate the missing matches, the ontologies need to be exposed to the expert in cleartext. Therefore, in our second framework, we try to limit the exposure of the ontologies only to the ontology-matching expert.

Preliminaries

Ontology mapping techniques can be classified into the following categories:

1. Word Similarity Based: In this case the concepts across ontologies are matched using the similarity of the words that appear in the ontologies (Mitra, Wiederhold, Decker).
2. Structural Similarity Based: This set of algorithms use the structure of the ontologies to match the concepts in the ontologies. (Melnik, Garcia-Molina, Rahm), (Noy and Musen).
3. Instance Based: Concepts in ontologies are matched using the similarity of their instances. Among the instance-based algorithms, we can further sub-classify them into two types:
   a. *Opaque Matching*: In this case, the matching does not depend upon the values of the instances but on the statistical properties, like distribution, entropy, mutual information etc. of the instances of a concept (Kang and Naughton).
   b. *Pattern-based Matching*: In this case, the algorithm identifies patterns in the values of the instances and uses similar patterns in their values to indicate that two concepts are similar.
4. Inference Based*:* The semantics of concepts in ontologies are expressed as rules using a logical language (say, the Web Ontology Language, OWL). Using an inference engine and these ontology rules, concepts across ontologies can be matched.

There are also algorithms that use hybrid or multiple strategies (Doan et al.).

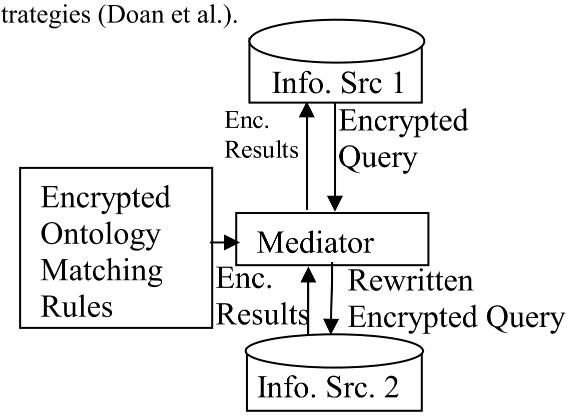

Figure 1. Privacy –preserving Interoperation System Architecture Using Private Key

Privacy-preserving Automated Ontology Matching

In this scenario, we assume that the organizations, say A and B, seeking to enable interoperation have a symmetric private key, K_{A-B}. The queries originating from both A and B, posed to the mediated system (as shown in Figure 1) are encrypted using the private key $K_{A-B.}$ The ontology matching rules used by the mediator are also encrypted using the same key.

We look at the ontology matching algorithms used to generate the encrypted ontology matching rules. As shown in Figure 2, the input to the ontology matcher, the source ontologies corresponding to the information sources for both A and B, are encrypted using K_{A-B}. The automated ontology matcher operates on the encrypted ontologies to match concepts across the ontologies.

Several ontology matching algorithms use dictionaries, thesauri or corpuses of documents to identify matching concepts. In order for these algorithms to work, the dictionaries, thesauri, or corpuses should also be encrypted using the same key, $K_{A-B.}$

Structure-based ontology matching techniques work fine even if the terms and relationships of the ontologies are encrypted because either they do not depend on the terms or relationships or even if they do, as long as the same labels are similarly encrypted, these algorithms are unaffected.

Instance-based matching algorithms that are opaque work fine without any modification because the statistical properties, like distribution, frequency, entropy, mutual information, of the instance values are not changed by encrypting it, however, pattern-based matching algorithms do not work because in most encryption systems destroy the patterns in the instance values.

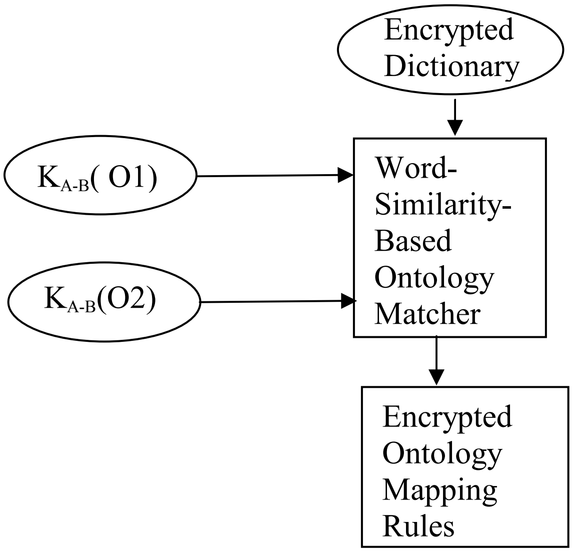
Encrypted Dictionary
K_{A-B}(O1)
K_{A-B}(O2)
Word-Similarity-Based Ontology Matcher
Encrypted Ontology Mapping Rules

Figure2. Privacy-preserving Automated Ontology Mapping

Semi-automated Ontology Matching Framework

The framework shown above has two important drawbacks:

1. It assumes that the process of ontology matching can be totally automated. Note that in the process outlined above, the ontology matcher has access to only encrypted ontologies and cannot decrypt the ontologies. Typically, human experts cannot match encrypted ontologies. Even if they do, matching ontologies without knowing the semantics of the concepts (because they are encrypted and thus their semantics is undecipherable) will not result in very accurate match generation.
2. Because the two organizations use a shared symmetric private key, each organization can observe the communication of the other with the mediator and obtain the other organization's ontology. In cases, where the organizations do not want to share their ontologies even with the organization they are sharing information with, such an arrangement is not acceptable.

In order to remedy the above-mentioned drawbacks, we offer the following interoperation and ontology-matching framework.

Interoperation Framework

Because the mediator cannot be trusted, the queries are encrypted and sent to the mediator. In this framework, each organization has a *unique* secret private key that it uses to encrypt the queries and ontologies. The mediator uses encrypted ontology-matching rules. For example, an encrypted ontology matching rule may be

((Y InstanceOf K2(O2.Car)) &
(Z Equals Y.K2(O2.Price)) &
(Z > 40,000)
=> (Y InstanceOf K1.(O1.LuxuryCar)) (R1)

The rule above indicates that if Y is an instance of car(O2.Car), Z is the the price(O2.Price) of Y, and Z is greater than 40,000, then Y is also an instance of luxury car. The keys K2 and K1 encrypt terms in O2, O1 respectively.

Using such an encrypted rule, the mediator can rewrite a given query, e.g., (?X InstanceOf K1(O1.LuxuryCar)) that asks for all instances of luxury-car(O1.LuxuryCar), by substituting the left-handside of (R1) for the query. Note that all the ontology terms in the query and the ontology-matching rule are encrypted and thus the mediator does not have access to those terms.

Privacy-preserving Semi-automated Ontology Matching

If we intend to use a human expert in the process of ontology matching, the human expert must have access to the ontologies in cleartext because encrypted labels will make no sense to the expert. In this scenario, as shown in Figure 3, each organization encrypts the ontologies using a session key that it shares with the expert (ontology matcher). Upon receiving the ontologies to be matched, the expert decrypts the encrypted ontologies using the session key. Each organization also has a public key that it has publicized via a certifying authority. The certifying authority serves as the trusted intermediary between the expert and the organizations. The expert (using a semi-automatic ontology matcher) matches the two ontologies and then creates a set of ontology matching rules similar to the rule shown in the example above. Let us say the source ontologies being matched are O1 and O2. Terms appearing in a ontology matching rule and O1 are encrypted using the public key (K1) of the first organization and terms appearing in the ontology matching rule and O2 are encrypted using the public key (K2) of the second organization (Figure 4). The mediator rewrites a query obtained from one organization, say Org1, encrypted using its key K1, to a query where all terms are from another organization's, say Org2, ontology, encrypted using its key K2 using the ontology mapping rules.

Related Work

Clifton et al., have argued about the need for and highlighted issues in privacy-preserving data integration and sharing. Agarwal and Srikant have shown how to mine data while preserving privacy. However, to the best of our knowledge, there exists no prior research that shows how privacy-preserving ontology matching can be enabled. Our interoperation architectures have been influenced by existing works on access-control in information interoperation systems (Damiani et al., Dawson, Qian and Samarati, de Capitani di Vimercati and Samarati, Gong and Qian).

Though there does not exist work on privacy-preserving ontology matching, as discussed above, several existing works have provided algorithms for ontology matching (Melnik, Garcia-Molina, and Rahm, Noy and Musen, 2001, Doan, et al., Hovy, Euzenat and Volchev, Shvaiko, Giunchiglia, and Yatskevich, Mitra, Wiederhold, and Decker, and Noy and Musen, Prasad, et al).

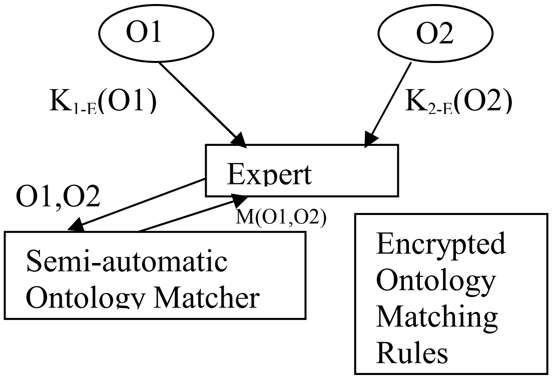
O1
O2
K_{1-E}(O1)
K_{2-E}(O2)
Expert
O1,O2
M(O1,O2)
Semi-automatic Ontology Matcher
Encrypted Ontology Matching Rules

Figure 3. Privacy-preserving Semi-automatic Ontology Matching

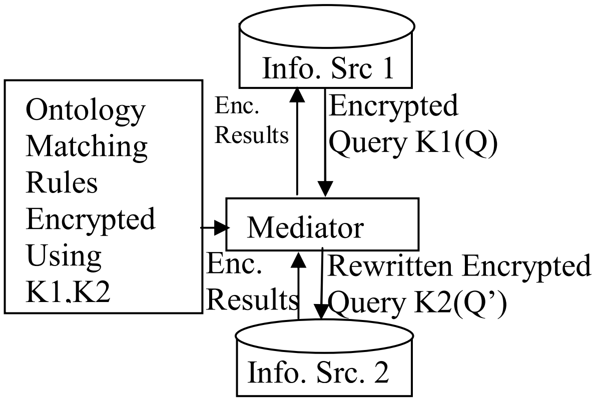
Info. Src 1
Enc. Results
Encrypted Query K1(Q)
Ontology Matching Rules Encrypted Using K1.K2
Mediator
Enc. Results
Rewritten Encrypted Query K2(Q')
Info. Src. 2

Figure 4: Privacy-preserving Interoperation Using Public Keys

Conclusion

Maintaining privacy in interoperation systems is becoming increasingly important. Ontology matching is the primary means of resolving semantic heterogeneity. Ontology matching helps establish semantic correspondence rules that are used for query rewriting and translation in interoperation systems. For information systems that want maximum privacy, the privacy of their ontologies must be maintained. In this paper, we describe two frameworks for privacy-preserving interoperation and show how we can implement privacy-preserving ontology matching.

References

Agrawal, R. and Srikant, R. Privacy-Preserving Data Mining. *In Proc. of the ACM SIGMOD, 2000.*

Clifton, C., Doan, A., Elmagarmid, A., Kantarcioglu, M., Schadow, G., Suciu, D., and Vaidya, J. Privacy Preserving Data Integration and Sharing, *Proc. of the 9th Int. Workshop on Data Mining and Knowledge Discovery, 2004 (DMKD-04)*

Damiani, E., De Capitani di Vimercati, S., Fugazza, C., and Samarati, P. Extending Policy Languages to the Semantic Web. *ICWE* 2004 330-343

Dawson, S., Qian, S., Samarati, P. Providing Security and Interoperation of Heterogeneous Systems. *Distributed and Parallel Databases*, vol. 8, no. 1, Jan. 2000, 119-145.

De Capitani di Vimercati, S., and Samarati, P. Authorization Specification and Enforcement in Federated Database Systems. *Journal of Computer Security*, vol. 5, no. 2, 1997, 155-188.

Doan, A., Madhavan, J., Domingos, P., and Halevy, A. Learning to map between ontologies on the semantic web. *In The Eleventh International WWW Conference*, Hawaii, US, 2002.

Euzenat, J., and Valtchev, P. Similarity-based ontology alignment in OWL-Lite. *In The 16th European Conference on Artificial Intelligence (ECAI-04),* Valencia, Spain, 2004.

Gong, L. and Qian, X. The Complexity and Composability of Secure Interoperation. *IEEE Symp. Security and Privacy*, (Oakland, CA, USA. 1994).

Hovy, E. Combining and standardizing largescale, practical ontologies for machine translation and other uses. *In The First International Conference on Language Resources and Evaluation (LREC)*, pages 535–542, Granada, Spain, 1998.

Kang, J., Naughton, J. F.: On Schema Matching with Opaque Column Names and Data Values. Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), San Diego, California, June 2003.

Melnik, S., Garcia-Molina, H. and Rahm, E. Similarityflooding: A versatile graph matching algorithm and its application to schema matching. *In 18th International Conference on Data Engineering (ICDE-2002)*, San Jose, California, 2002. IEEE Computing Society.

Mitra, P. and Wiederhold, G. Resolving terminological heterogeneity in ontologies. *In Workshop on Ontologies and Semantic Interoperability at the 15th European Conference on Artificial Intelligence (ECAI)*, July 2002.

Mitra, P., Wiederhold, W., and Decker, S. A scalable framework for interoperation of information sources. *In The 1st International Semantic Web Working Symposium (SWWS'01),* Stanford University, Stanford, CA, 2001.

Noy, N.F., and Musen, M.A. Anchor-PROMPT: Using non-local context for semantic matching. *In Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, Seattle, WA, 2001.

Noy, N. F., and Musen, M. A.. The PROMPT suite: Interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, 59(6):983–1024, 2003.

Prasad, S., Peng, Y., and Finin, T. A tool for mapping between two ontologies using explicit information. *In AAMAS 2002 Workshop on Ontologies and Agent Systems,* Bologna, Italy, 2002.

Shvaiko, P., Giunchiglia, F. and Yatskevich, M. S-Match: an Algorithm and an Implementation of Semantic Matching, *Proc. of the 1st European Semantic Web Symposium, 2004 (ESWS-04).*