RONTO: RELATIONAL TO ONTOLOGY SCHEMA MATCHING

Petros Papapanagiotou, Polyxeni Katsiouli, Vassileios Tsetsos, Christos Anagnostopoulos and Stathes Hadjiefthymiades

University of Athens, Department of Informatics and Telecommunications,
Pervasive Computing Research Group
Panepistimiopolis, Ilissia, 15784, Athens, Greece

P.Papapanagiotou@sms.ed.ac.uk, {polina, b.tsetsos, bleu, shadj}@di.uoa.gr

http://p-comp.di.uoa.gr

***Abstract:* The population of ontologies with real instances still remains a major practical issue for the Semantic Web community. Significant progress towards a solution for this issue can be achieved through the migration of data stored in existing relational databases. In this paper we present a schema matching methodology and its implementation, necessary for further data migration.**

1. Introduction

Semantic Web (SW) [4] is already in its implementation phase, with ontologies playing a core modeling role. Great progress has already been achieved in many SW areas, such as ontological engineering, Description Logics (DL) reasoning, and query languages. However, an important problem still remains unsolved: *lack of real semantically-annotated data*. With the proposed system, called RONTO, we try to address this problem through a) the schema matching between relational schemata and SW ontologies, and, b) the population of SW ontologies with data from relational databases. The rational behind our approach is based on the fact that a lot of information in the Web is stored in relational databases, which form the so-called Deep Web [3]. In the following sections we briefly describe the RONTO system design and prototype implementation, focusing on the schema matching processes.

2. Related Work

Schema matching is a research field that has attracted the interest of the data and knowledge engineering community. Most researchers study schema matching in a specific context (e.g., relational to object-oriented, relational to XML schema). Some researchers have also tried to generalize the matching process and have proposed generic algorithms for schema mapping. Concerning the relational to ontology case, KAON Reverse [9] is among the first semi-automatic tools for schema matching and data migration. It adopts a reverse engineering approach and the schema mapping is based on fixed rules, which are defined manually by users. COMA++ [2] is another matching tool with a graphical user interface. The main characteristic of COMA++ is the fact that it combines different matchmaking algorithms. COMA++ provides also the user with the ability to compose, merge and reuse existing mappings. Finally, MapOnto [11] is an ongoing project, which establishes semantic mappings between database schemata and ontologies as well as between different database schemata.

3. Schema Matching Methodology

3.1 Definitions

Before discussing the adopted methodology we should make some assumptions for the main elements involved in RONTO. Firstly, we assume that the source schema is a relational database schema, RDB, deployed on a typical commercial relational database management system. We also assume that the conceptual schema of the target ontology (ONT) is expressed in a DL language, due to the popularity of DLs in the SW community.

Moreover, in order to better describe the presented methodology, several intermediate modeling elements are introduced:

Definition 1. A *Candidate Concept* for an ontology concept *c*, CC_c , may be (i) an RDB relation, or (ii) an RDB view or (iii) a combination of them, which is *structurally* and "*semantically* similar" to the concept *c* of the target ontology.

Definition 2. A *Candidate Datatype-Property* for a datatype-property *p*, CDP_p, is an attribute of an RDB relation[2], which has the *same* (or a *compatible*) *data type,* and is "*semantically* similar" to the datatype-property *p* of the target ontology. Similarly we can define the *Candidate Object-Property* for an object-property *p*, $COP_{p.}$

Definition 3. A *Candidate Concept Set*, CCS_C, for an ontology concept *c* is the set of all CCs that can be computed for the concept *c*. Similarly, Candidate Datatype-property Sets (CDPS) and Candidate Object-Property Sets (COPS) are defined. Each element *e* of such sets is associated with a degree of similarity, *sim(e, r)*, where *r* is an element belonging to the target ontology. The similarity *threshold*, (i.e., the minimum acceptable similarity value) depends on the user.

As already stated, the RONTO methodology for schema matching is heavily based on different types of similarity (linguistic, semantic similarity and data type compatibility) and exploits a variety of similarity measures in order to effectively perform the schema matching. Linguistic similarity measures compare the schema elements based on the lexicographic characteristics of their names/labels. Semantic similarity measures are used for schema element names that are valid words. In order to compute such similarity,

[2] We assume columns that are not foreign keys.

techniques like those described in [7] are used. The compound similarity between two schema elements, $a \in$ RDB and $b \in$ ONT, is the weighted sum of the aforementioned similarity measures.

3.2 Matching Steps

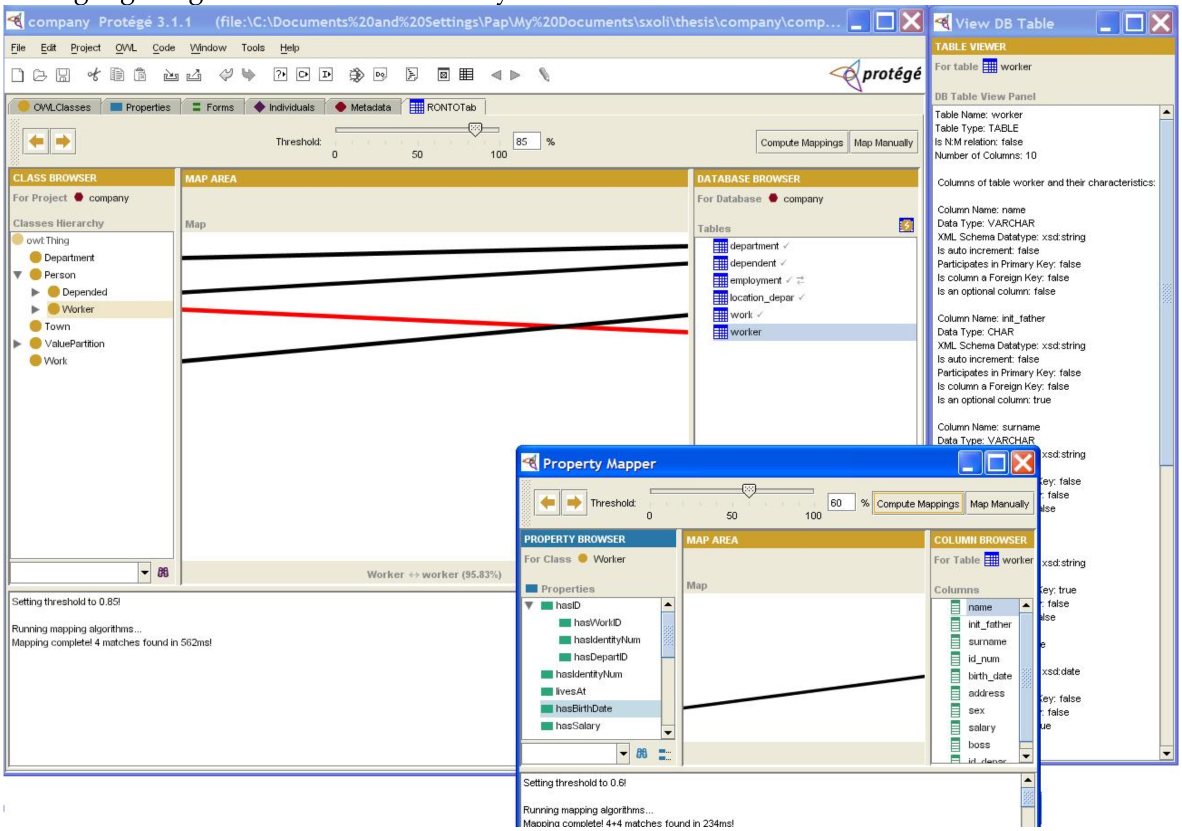
In order to achieve the schema matching and data migration procedures, we have designed a complete methodology which is based on similarity measures in order to assess mappings. This methodology is composed of the following algorithms:

- **Tables to Concepts Mapping.** *We find all the CCs for each concept c ∈ ONT. Note that the database tables representing N:M relationships between two different relations are excluded from this mapping phase.*
- **Attributes to Datatype-properties Mapping.** *The methodology proceeds with the computation of the mappings between the relation attributes and the datatype-properties of the ontology. Foreign keys are excluded from this step. In this step, we consider not only the (linguistic and semantic) similarity between the labels of the elements, but also the data type compatibility between the RDB attributes and the range of each datatype-property.*
- **Foreign Keys to Object-properties Mapping.** *According to the OWL-DL [1] language, an object-property expresses a binary relationship between two concepts of the ontology. In relational databases, relationships among tables are expressed through referential constraints, represented by foreign keys. The present process defines mappings between such database elements and the object-properties of the ontology.*
- ***N:M Relations to Object-properties Mapping***. Except from the referential constraints, there are also database relations which represent binary relationships between two tables. Such relations may constitute COPs for the object-properties of the ontology.
- ***Joined Tables to Concepts Mapping***. There are some cases in which the information content carried by one concept is distributed in more than one relations of the database (i.e., joins). Therefore, RONTO computes all the possible joins between the database relations. Next, it applies a two-step algorithm. In the first step, the algorithm clusters the database relations which have an attribute similar to a datatype-property of a specific concept c. Moreover, it computes all possible joins between different relations from different sets (i.e., CCS_c). During the second step, the algorithm eliminates, for each object-property p of concept c with range R, all the CCs from the CCS_c which do not have a foreign key referencing the primary key of a CC which belongs to the CCS_R. The same rule is also applied to CCs from CCS_R, which do not contain a primary key referenced by a foreign key from a CC of CCS_c.
- ***Attributes to Object-properties Mapping.*** This step deals with cases where a database attribute can be mapped to an object-property of the ontology.

3.3 Implementation Details

RONTO is a tool developed as a Protégé plug-in, since Protégé [6] is currently the most popular open-source ontology editor with a very large community of users and developers. The mapping is performed under human supervision through a friendly graphical user interface. RONTO makes use of the Protégé OWL Plug-in API for handling ontologies in conjunction with a JDBC-based module for handling relational databases. The latter module extracts the database meta-data and presents it to the user in tree-like structures similar to the ones used by Protégé for presenting conceptual hierarchies. Users can request additional information about the database, including all the possible joins between the database relations. RONTO guides, in a step-by-step way, the users through the matching process. Starting from the Tables to Concepts mapping, RONTO visualizes all the automatically calculated matches and their respective similarity measures as lines connecting the two matching elements. The user can accept or reject the proposed mappings before proceeding to the next step. She can also manually map elements whose similarity has not been correctly detected by the system.

RONTO uses a variety of similarity techniques. Users can choose which of these techniques should be used for each mapping step and may also tweak the similarity threshold at every step. Thus, they have full control over the automated part of the schema matching procedure. The results produced by this prototype version of RONTO can be stored either in a proprietary or a D2R Map format. D2R Map [10] is a declarative language to describe mappings between relational database schemata and OWL/RDFS ontologies. We have performed a preliminary evaluation of the RONTO prototype with some artificial datasets and compared the results with some expert mappings, which were obtained by performing the task manually. The evaluation was based on metrics commonly uses in schema matching such as precision, recall and F-measure [5]. RONTO demonstrated high precision and recall values, especially in large schemata and schemata with elements having high degree of semantic similarity.



Threshold: 85 %

```
Setting threshold to 0.85!

Running mapping algorithms...
Mapping complete! 4 matches found in 562ms!
```

```
Table Name: worker
Table Type: TABLE
Is N:M relation: false
Number of Columns: 10

Columns of table worker and their characteristics:

Column Name: name
Data Type: VARCHAR
XML Schema Datatype: xsd:string
Is auto increment: false
Participates in Primary Key: false
Is column a Foreign Key: false
Is an optional column: false

Column Name: init_father
Data Type: CHAR
XML Schema Datatype: xsd:string
Is auto increment: false
Participates in Primary Key: false
Is column a Foreign Key: false
Is an optional column: true

Column Name: surname
Data Type: VARCHAR
```

Threshold: 60 %

```
Setting threshold to 0.6!

Running mapping algorithms...
Mapping complete! 4+4 matches found in 234ms!
```

Figure 1. Screenshot of the RONTO plug-in

4. Conclusion

We have briefly presented the design of a new tool for schema matching that addresses in a practical way the requirement for actual data in the Semantic Web. The main steps of the matching methodology were presented along with some implementation details. However, there are a lot of open issues in SW-related schema matching research, especially when the schemata are very diverse. For instance, it would be very useful if we could automatically identify n-ary database relationships and map them to ontological properties. This task becomes more challenging if we consider that there is no standard representation of n-ary relationships in databases and ontologies. We are currently working on the improvement of our matching methodology by taking into account the cardinality constraints that may exist in a conceptual schema. Finally, we try to build appropriate datasets in order to evaluate the performance and effectiveness of our approach and compare it to other existing approaches.

6. References

[1]. Antoniou, G., van Harmelen, F. (2004 April). A Semantic Web Primer, The MIT Press.

[2]. Aumueller, D., et al. (2005). Schema and Ontology Matching with COMA++, SIGMOD.

[3]. Bergman M. (2000). The Deep Web, Surfacing Hidden Value, The Journal of Electronic Publishing 7 (1).

[4]. Berners-Lee, T., et al. (2001, May). The semantic web, Scientific American.

[5]. Do, H., et al. (2002). Comparison of Schema Matching Evaluations, Proceedings GI-Workshop "Web and Databases"

[6]. Gennari, J., et al. (2002). The Evolution of Protege: An Environment for Knowledge-Based Systems Development. International Journal of Human-Computer Studies

[7]. Pedersen, T., et al. (2004). Wordnet::similarity – measuring the relatedness of concepts. Fifth Annual Meeting of the North American Chapter of the ACL (NAACL-04), Boston, MA.

[8]. Rahm, E., Bernstein, P.A. (2001). A Survey of Approaches to Automatic Schema Matching. VLDB Journal, 10(4):334-350.

[9]. Stojanovic, N., et al. (2002). A reverse engineering approach for migrating data-intensive web sites to the Semantic Web, In Proceedings of the Conference on Intelligent Information Processing, World Computer Congress, Montreal, Canada, Kluwer, Academic Publishers.

[10]. Bizer, C. (2003). D2Rmap- A Database to RDF Mapping Language, Budapest, Hungary, Poster at the 12th World Wide Web Conference.

[11]. Yuan, A., et al. (2005). Inferring Complex Semantic Mappings between Relational Tables and Ontologies from Simple Correspondences. In Proceedings of CoopIs/DOA/ODBASE.

	Petros Papapanagiotoy, Greece received his B.Sc. in Informatics from the Department of Informatics & Telecommunications at the University of Athens, Greece in 2006. Nowadays he is a M.Sc. student at University of Edinburgh. His research focuses on Semantic Web and knowledge representation and reasoning.
	Polyxeni Katsiouli, Greece received her B.Sc. in Informatics from the Department of Informatics & Telecommunications at the University of Athens, Greece in 2006. Nowadays she is a M.Sc. student at Athens University of Economics and Business. Her research focuses on Semantic Web, semantic hypermedia and ontology learning.
	Vassileios Tsetsos, Greece received his B.Sc. in Informatics from the Department of Informatics & Telecommunications at the University of Athens, Greece in 2003 and a M.Sc. in "Communication Systems and Data Networks" from the same department in 2005. Nowadays he is a Ph.D. student at the same department. His research focuses on Semantic Web services, ontological engineering and pervasive computing.
	Christos Anagnostopoulos, Greece received his B.Sc. in Informatics from the Department of Informatics & Telecommunications at the University of Athens, Greece in 2001 and a M.Sc. in "Advanced Information Systems" from the same department in 2003. Nowadays he is a Ph.D. student at the same department. His research focuses on context-/situation-awareness, contextual reasoning and pervasive computing.
	Stathes Hadjiefthymiades, Greece received his B.Sc. in Computer Science from the Department of Informatics at the University of Athens, Greece in 1993 and his M.Sc. in Advanced Information Systems from the same department in 1996. In 1999 he received his Ph.D. from the same department. In June 2002 he received a Joint Engineering-Economics M.Sc. degree from the National Technical University of Athens. Since December 2003 he is an Assistant Professor at the Dept. of Informatics and Telecommunications, University of Athens. His research interests are in the areas of wireless/mobile/pervasive computing and networked multimedia applications.