R2Sim: A Novel Semantic Similarity Measure for Matching between RDF Schemas

PHAM THI THU THUY, YOUNG-KOO LEE, SUNGYOUNG LEE
Department of Computer Engineering
Kyung Hee University
KOREA
tttpham@oslab.khu.ac.kr, yklee@khu.ac.kr, sylee@oslab.khu.ac.kr

Abstract

Information integration for distributed and heterogeneous data sources has recently gained great attention, and schema matching plays a prominent role in this process. This paper presents R2Sim, a novel method to automatic concept matching between RDF Schemas using semantic similarity measure which considers both description and neighborhood similarities. The description similarity is the combination of thee components: name, definition and data type, while the neighborhood similarity considers the root, parent and child factors. The key features of R2Sim are that it propose novel metric to measure the data type similarity and the *relaxation labeling*, a well-known constraint optimization technique used in computer vision and other fields, can be adapted to work efficiently in our method.

Keywords: RDF Schema; Semantics; Similarity measure, Ontology.

1 Introduction

RDF Schema (RDFS), an RDF vocabulary description Language, has been a widely-used language for representing data in the Web [1]. The number of RDFS increasing leads to the heterogeneity problem. The same concepts may be modeled differently by using different terms or placed in different positions in the concept hierarchy. Therefore, measuring the concept similarity in two RDFSs is the core for the success of information integration, or the RDFS matching.

Several approaches have been proposed to measure the concept similarity. Most of them mainly focused on concepts within a single ontology. There are a few publications which address the determining concept similarity from different ontologies [2-4]. However, most of these methods only consider the information which describes the concepts such as name, definition, and property. Further, the similarity values of some factors such as data type and definition are given by the users' judgment.

This paper presents R2Sim method that measures the semantic similarity between concepts from different RDF Schemas. The semantics of the concepts are implied in name, their description and their relationship with other concepts in the schema tree. Taking into account the description and the neighborhood similarity of two RDF Schemas, a relaxation labeling algorithm is developed to optimize the concept matching among schemas.

The remainder of the paper is organized as follows. A motivating example for this paper is presented in Section 2. Section 3 describes our R2Sim approach. Finally Section 4 concludes the paper.

2 Motivating Example

In order to motivate our method, we use two real-world RDF Schemas. The first schema is taken from the RDF Primer [1], while the second one is extracted from the book [5]. They are presented in Fig.1 and Fig.2, respectively.

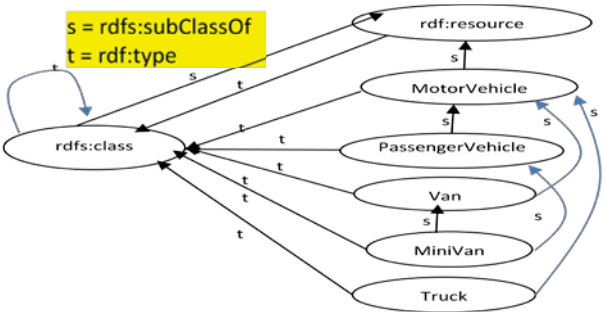

Fig. 1. The MotorVehicle RDF Schema

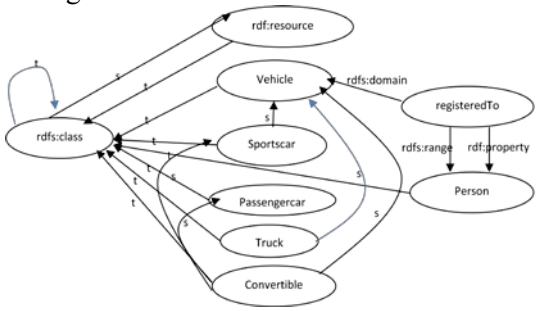

Fig. 2. The Vehicle RDF Schema

Acknowledgment: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (2011-0020499).

Each concept in figures is associated with a set of primitives, such as *rdf:id, rdfs:range, rdfs:subClassOf*, etc. Each primitive brings a piece of knowledge to the whole meaning of the concept. The motivation is to find the most suitable matching from each concept in Fig.1 to one concept in Fig.2.

3 R2Sim

To match concepts between two RDF Schemas, we need a full computation of concept similarity. Our R2Sim measuring is determined by the following equation:

$$R2Sim(C_1, C_2) = \frac{\alpha_1 * DcSim(C_1, C_2) + \alpha_2 * NbSim(C_1, C_2)}{\alpha_1 + \alpha_2} \quad (1)$$

where C_1 and C_2 are the concepts of the first and the second RDF Schema, respectively; α_1 and α_2 are the weight parameters between 0 and 1; *DcSim* is the Description similarity; *NbSim* is the neighborhood similarity.

3.1 Description Similarity

The description similarity between two concepts C_1 in $RDFS_1$ and C2 in RDFS2 comprises three factors: Name similarity (*NSim*), definition similarity (*DfSim*), and Data type similarity (*DtSim*). It is determined by following equation (2):

$$DcSim(C_1, C_2) = \frac{\beta_1 * NSim(C_1, C_2) + \beta_2 * DfSim(C_1, C_2) + \beta_3 * DtSim(C_1, C_2)}{\beta_1 + \beta_2 + \beta_3} \quad (2)$$

where β_1, β_2 and β_3 are the weight parameters between 0 and 1; *DfSim* is calculated based on the similarity of the definition given for each concept.

For computing the name similarity, if the concept is declared as the set of word of a short form of some words, the normalization and tokenization steps are required. The name similarity between two concepts C_1 and C_2 is extracted from below equation (3):

$$NSim(C_1, C_2) = \frac{n_{C_1 \cap C_2}}{\max(n_{C_1}, n_{C_2})} \quad (3)$$

where $n_{C_1 \cap C_2}$ is the number of matching characters between elements C_1 and C_2; *max* is the maximum value; n_{C_1} and n_{C_2} are the lengths of the elements C_1 and C_2, respectively.

Since most of RDF Schema's data types are similar to those of XML Schema, based on the characteristics of each data type [6], we define the metric for measuring the similarity among the data types (4):

$$DtSim(C_1, C_2) = \frac{\sum_i \left| \{ cf_i \mid C_1[cf_i] = C_2[cf_i], 1 \le i \le n_{cf} \} \right|}{n_{cf}} \quad (4)$$

where *cf* is the list of the constraining facets described in [6], n_{cf} is the total number of the constraining facets.

3.2 Neighborhood Similarity

The neighborhood similarity (*NbSim*) between two concepts C_1 in $RDFS_1$ and C_2 in $RDFS_2$ is the combination of three components: Root node similarity (*RSim*), parent similarity (*PSim*), and child similarity (*ChSim*). It is computed based on the following equation (5):

$$NbSim(C_1, C_2) = \frac{\varepsilon_1 * RSim(C_1, C_2) + \varepsilon_2 * PSim(C_1, C_2) + \varepsilon_3 * ChSim(C_1, C_2)}{\varepsilon_1 + \varepsilon_2 + \varepsilon_3} \quad (5)$$

where ε1, ε2 and ε3 are the weight parameters between 0 and 1; *RSim* is computed based on the distance similarity from each concept to the root node; *PSim* and *ChSim* are the average functions of the equal directed super-concepts and child-concepts per the total number of super-concepts and child-concepts, respectively.

4 Conclusion

The approach to the concept matching between RDF Schemas has been presented. The similarity between for the concepts between RDF Schemas is calculated based on the concept descriptions, including concept name, concept definitions, its data type, and the neighboring concepts. Based on the similarity measure of two concepts, the relaxation modeling is applied to find the most suitable matching between concepts of two RDF schemas.

References

[1] Frank Manola, Eric Miller, *W3C*, 2004, http://www.w3.org/TR/rdf-primer/#rdfschema

[2]Doan A. H., Madhavan J., Domingos P., Ontologies Matching: A Machine Learning Approach, *Handbook on Ontologies in Inf. Systems*, Springer-Velag, 2003.

[3] Ehrig M., Sure Y., Ontology Mapping – an integrated approach, *1st European Semantic web Symposium*, Greece, 2004.

[4] Oundhankar S., K. Verma, Sivashanugam K., Discovery of web services in a Multi-Ontologies and Federated Registry Environment, *International Journal of Web Services Research,* 1, 3, 2005.

[5] Rodriguez M. A., Egenhfer M. J., Determining Semantic Similarity among entity classes from different ontologies, *IEEE Transactions of Knowledge and Data Engineering* 15, 2, 442 1041-4347, 2003.

[5] Ronald M., Thomas H., Rene P., Enterprise Knowledge Infrastructures, 2nd edition, Springer, 2009.

[6] D Vint Productions, "XML Schema - Data Types Quick Reference", http://www.xml.dvint.com, 2003.