Generic Similarity Detection in Ontologies with the SOQA-SimPack Toolkit[*]

Patrick Ziegler Christoph Kiefer Christoph Sturm
Klaus R. Dittrich Abraham Bernstein
Department of Informatics, University of Zurich
{pziegler, kiefer, sturm, dittrich, bernstein}@ifi.unizh.ch

ABSTRACT

Ontologies are increasingly used to represent the intended real-world semantics of data and services in information systems. Unfortunately, different data sources often do not relate to the same ontologies when describing their semantics. Consequently, it is desirable to have information about the similarity between ontology concepts for ontology alignment and integration. In this demo, we present the SOQA-SimPack Toolkit (SST), an ontology language independent Java API that enables generic similarity detection and visualization in ontologies. We demonstrate SST's usefulness with the SOQA-SimPack Toolkit Browser that allows users to graphically perform similarity calculations in ontologies.

1. INTRODUCTION

In current information systems, ontologies are increasingly used to explicitly represent the intended real-world semantics of data and services. Ontologies provide a means to overcome heterogeneity by providing explicit, formal descriptions of concepts and their relationships which exist in a certain universe of discourse, together with a shared vocabulary to refer to these concepts. Based on agreed ontological domain semantics, the danger of semantic heterogeneity can be reduced. Ontologies can, for instance, be applied in the area of data integration for data content explication to ensure semantic interoperability between data sources.

Unfortunately, different data sources often do not relate to the same ontologies when describing their semantics. That is, schema elements can be linked to concepts of different ontologies in order to explicitly express their intended meaning. This complicates the task of finding semantically equivalent schema elements since at first, semantic relationships between the concepts to which the schema elements are linked to have to be detected. Consequently, it is desirable to have information about the similarity between ontological concepts. In addition to schema integration, such similarity information can be useful for many applications, such as ontology alignment and integration, Semantic Web (service) discovery, data clustering and mining, semantic interoperability in virtual organizations, and semantics-aware universal data management.

The task of detecting similarities in ontologies is aggravated by the fact that a large number of ontology languages is available to specify ontologies. Besides traditional ontology languages, such as Ontolingua [3] or PowerLoom[1], there is a notable number of ontology languages for the Semantic Web, such as SHOE[2], DAML[3], or OWL[4]. That is, data semantics are often described with respect to ontologies that are represented in various ontology languages. In consequence, mechanisms for effective similarity detection in ontologies must be capable of coping with heterogeneity caused by the use of different ontology languages. Additionally, it is desirable that different similarity measures can be employed so that different approaches to identify similarities among concepts in ontologies can be reflected.

For instance, assume that in an example scenario, a developer of an integrated university information system is looking for semantically similar elements from database schemas which relate to the following ontologies to describe their semantics: (1) the Lehigh University Benchmark Ontology[5] that is represented in OWL, (2) the PowerLoom Course Ontology[6] developed in the SIRUP project [6], (3) the DAML University Ontology[7] from the University of Maryland, (4) the Semantic Web for Research Communities (SWRC) Ontology[8] modeled in OWL, and (5) the Suggested Upper Merged Ontology (SUMO)[9], which is also an OWL ontology. Assume further that there are schema elements linked to all of the 966 concepts which these five ontologies comprise. Unless suitable tools are available, identifying semantically related schema elements in this set of concepts and visualizing the similarities appropriately definitely turns out to be time-consuming and labor-intensive.

In this demonstration, we present the SOQA-SimPack

[*] `http://www.ifi.unizh.ch/dbtg/Projects/SIRUP/SST/`

[1] `http://www.isi.edu/isd/LOOM/PowerLoom/`

[2] `http://www.cs.umd.edu/projects/plus/SHOE/`

[3] `http://www.daml.org`

[4] `http://www.w3.org/2004/OWL/`

[5] `http://www.lehigh.edu/~zhp2/univ-bench.owl`

[6] `http://www.ifi.unizh.ch/dbtg/Projects/SIRUP/ontologies/course.ploom`

[7] `http://www.cs.umd.edu/projects/plus/DAML/onts/univ1.0.daml`

[8] `http://www.ontoware.org/projects/swrc/`

[9] `http://reliant.teknowledge.com/DAML/SUMO.owl`

SIGMOD 2006, June 27–29, 2006, Chicago, Illinois, USA.
Copyright 2006 ACM 1-59593-256-9/06/0006 ...$5.00.

Toolkit (SST), an ontology language independent Java API that enables generic similarity detection and visualization in ontologies. SST's main goal is to define a Java API for calculating and visualizing similarities in ontologies for a broad range of ontology languages. Considering the fact that different data sources often do not relate to the same ontologies, we support the calculation of similarities not only within a given ontology, but also between concepts of *different* ontologies. For these calculations, SST provides a generic and extensible library of ontological similarity measures capable of capturing a variety of notions of "similarity". Note that we do not focus on immediate ontology integration. Instead, we strive for similarity detection among different pre-existing ontologies, which are separately used to explicitly state real-world semantics as intended in a particular setting.

2. OVERVIEW OF THE SOQA-SIMPACK TOOLKIT

The SOQA-SimPack Toolkit (SST) [7] is an ontology language independent Java API that enables generic similarity detection and visualization in ontologies. The two foundations of SST are:

- The SIRUP Ontology Query API (SOQA) [8], an ontology language independent Java API for query access to ontological metadata and data. SOQA provides unified access to ontologies according to the SOQA Ontology Meta Model [8] that represents modeling capabilities that are typically supported by ontology languages to describe ontologies and their components; that is, concepts, attributes, methods, relationships, instances, and ontological metadata.[10]
- SimPack [2], a generic and extensible Java library of similarity measures adapted for the use in ontologies. The spectrum of currently provided similarity measures ranges from vector-based measures over string-, full-text-, tree-, and graph distance-based measures to information theory-based measures (see [7, 2] for more information).[11]

Simply stated, SST accesses data concerning concepts to be compared through SOQA; this data is then taken as an input for the similarity measures provided by SimPack. Thus, SST offers ontology language independent similarity calculation services based on the uniform view of ontological content as provided by the SOQA Ontology Meta Model. SST services that have already been implemented include:

- Similarity calculation between two concepts according to a single similarity measure or a list of measures.
- Similarity calculation between a concept and a set of concepts according to a single or a list of similarity measures. As for *all* sets of concepts provided as parameters in SST, this set of concepts can either be a freely composed list of concepts or all concepts from an ontology taxonomy (sub)tree.
- Retrieval of the k most similar concepts of a set of concepts for a given concept according to a single or a list of similarity measures.
- Retrieval of the k most similar concept pairs (c_1, c_2) from two given sets of concepts C_1 and C_2 according to a single similarity measure ($c_1 \in C_1$ and $c_2 \in C_2$).
- Retrieval of the k most *dis*similar concepts of a set of concepts for a given concept according to a single or a list of similarity measures.

Note that for all calculations provided by SST, the concepts involved can be from *any* ontology that is connected through SOQA.[12] That is, not only is it possible to calculate similarities between concepts from a single ontology (for example, Student and Employee from the DAML University Ontology) with a given set of SimPack measures, but also can concepts from different ontologies be used in the very same similarity calculation (for example, Student from the PowerLoom Course Ontology can be compared with Researcher from WordNet [5]). For all SST computations, the results can be output textually (floating point values or sets of concept names, depending on the service). Alternatively, calculation results can automatically be visualized and returned by SST as a chart.

Using concepts from different ontologies in the same similarity calculation is enabled by the fact that in SST, all ontologies are incorporated into a single ontology tree. That is, the root concepts of the available ontologies (e.g., owl:Thing) are direct subconcepts of a so-called Super_Thing root concept. This makes it possible that, for instance, not only vector- and text-based similarity measures, but also graph based measures that need a contiguous, traversable path between the concepts can be applied to concepts in SST.

3. DEMONSTRATION HIGHLIGHTS

To demonstrate the capabilities of SST, we assume the example scenario as presented in the introduction of this paper: a developer of an integrated university information system is looking for semantically similar elements from database schemas that relate to 966 concepts from five ontologies represented in three different ontology languages. We demonstrate SST's usefulness with the SOQA-SimPack Toolkit Browser that allows users to graphically perform similarity calculations and visualizations in ontologies based on SST services.

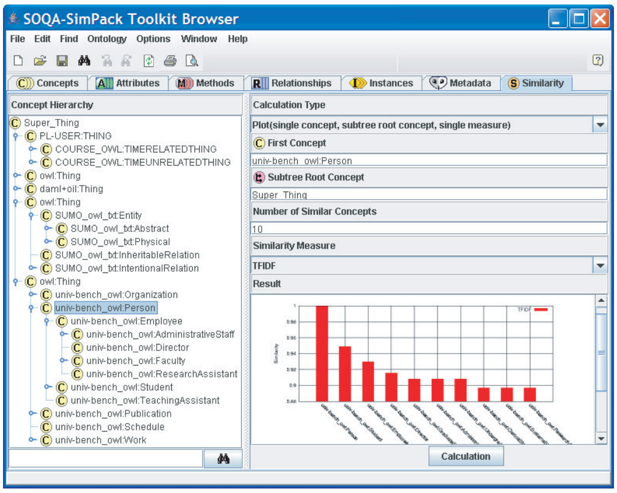
The SOQA-SimPack Toolkit Browser is an extension of the SOQA Browser [8] enabling users to inspect the contents of ontologies independently of the particular ontology language (i.e., according to the SOQA Ontology Meta Model). Based on the unified view of ontologies it provides, the SST Browser can be used to quickly survey concepts and their attributes, methods, relationships, and instances that are defined in ontologies as well as metadata (author, version, ontology language name, etc.) concerning the ontology itself.

In addition, the SST Browser provides an interface to all the methods of the SOQA-SimPack Toolkit through its Similarity Tab (see Figure 1 and 2). That is, it is a tool for performing language independent similarity calculations in ontologies and for result visualization. In the Similarity Tab,

[10] http://www.ifi.unizh.ch/dbtg/Projects/SIRUP/SOQA/

[11] http://www.ifi.unizh.ch/ddis/simpack.html

[12] Generally, this is every ontology that can be represented in an ontology language. In fact, it is every ontology that is represented in a language for which a SOQA wrapper is available (currently: OWL, DAML, PowerLoom, and WordNet [5]).



Calculation Type: Plot(single concept, subtree root concept, single measure)

First Concept: univ-bench_owl:Person

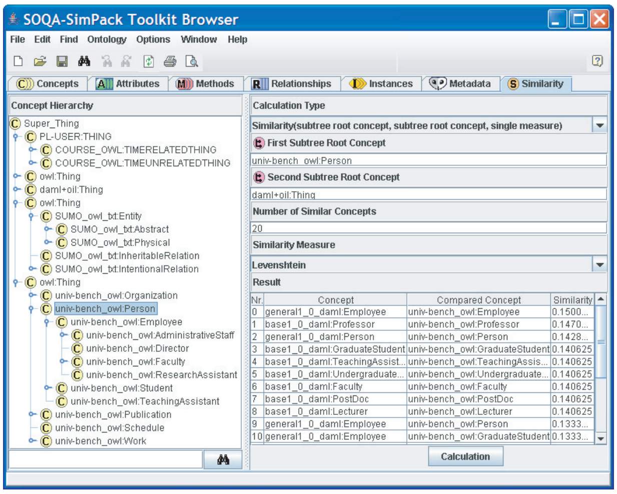
Subtree Root Concept: Super_Thing

Number of Similar Concepts: 10

Similarity Measure: TFIDF

Result

Figure 1: Calculating the most similar concepts for univ-bench_owl:Person compared with all concepts from all five ontologies with the TFIDF measure [1]



Calculation Type: Similarity(subtree root concept, subtree root concept, single measure)

First Subtree Root Concept: univ-bench_owl:Person

Second Subtree Root Concept: daml+oil:Thing

Number of Similar Concepts: 20

Similarity Measure: Levenshtein

Result

Nr.	Concept	Compared Concept	Similarity
0	general1_0_daml:Employee	univ-bench_owl:Employee	0.1500...
1	base1_0_daml:Professor	univ-bench_owl:Professor	0.1470...
2	general1_0_daml:Person	univ-bench_owl:Person	0.1428...
3	base1_0_daml:GraduateStudent	univ-bench_owl:GraduateStudent	0.140625
4	base1_0_daml:TeachingAssist...	univ-bench_owl:TeachingAssis...	0.140625
5	base1_0_daml:Undergraduate...	univ-bench_owl:Undergraduate...	0.140625
6	base1_0_daml:Faculty	univ-bench_owl:Faculty	0.140625
7	base1_0_daml:PostDoc	univ-bench_owl:PostDoc	0.140625
8	base1_0_daml:Lecturer	univ-bench_owl:Lecturer	0.140625
9	general1_0_daml:Employee	univ-bench_owl:Person	0.1333...
10	general1_0_daml:Employee	univ-bench_owl:GraduateStudent	0.1333...

Figure 2: Calculating the most similar concept pairs (c_1, c_2) for two ontology subtrees S_1, S_2 with the Levenshtein measure [4] $(c_1 \in S_1,\ c_2 \in S_2)$

users can select the similarity service to be run — for example, producing a graphical representation of the similarity calculation between two concepts according to the Levenshtein measure [4]. Depending on the selected service, appropriate fields are automatically inserted into the Similarity Tab so that all necessary input values can be entered; here, concept names can directly be mouse-dragged from the Concept Hierarchy view and dropped into the respective input field. In the end, the calculated results are shown in tabular or graphical form, depending on the selected service.

In the demo, we first employ the SOQA-SimPack Toolkit Browser to quickly provide a unified overview of the five ontologies represented in PowerLoom, OWL, and DAML respectively. Subsequently, we present the Similarity Tab and demonstrate similarity calculations, for instance, the k most similar concepts for univ-bench_owl:Person from all five ontologies according to the TFIDF measure (see Figure 1). We demonstrate similarity calculations in real-time and show how the results of these calculations can be presented as numerical values, textual lists (of concept names), or visualized in charts.

For our schema integration scenario of 966 concepts to which database schema elements are linked to, we particularly demonstrate how SST can be used to find the k most similar concepts for a given one from a set of concepts according to a variety of similarity measures. In addition, we illustrate that SST can effectively help to identify the most promising candidates for integration by retrieving the k most similar concept pairs from a given set of concepts (see Figure 2). Hence, we show how SST enables users to find semantically equivalent and related schema elements based on ontological similarity calculations. Analogously, similarity information, as provided by SST, can be employed in other areas, such as ontology alignment and integration, semantic interoperability in virtual organizations, and semantics-aware universal data management.

Contrasting a conventional scenario where several ontology-language specific tools have to be employed for ontology access, we illustrate that the developer who takes advantage of SST does not have to cope with different ontology representation languages in use. Based on the unified view of ontologies as provided by the SOQA Ontology Meta Model, it is interactively shown how our developer can generically apply a rich and extensible set of SimPack similarity measures for similarity detection through the services offered by the SOQA-SimPack Toolkit. Thus, we demonstrate how similarity detection in ontologies is facilitated and leveraged through SST and its browser.

4. REFERENCES

[1] R. Baeza-Yates and B. d. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

[2] A. Bernstein, E. Kaufmann, C. Kiefer, and C. Bürki. SimPack: A Generic Java Library for Similarity Measures in Ontologies. Technical report, University of Zurich, Department of Informatics, 2005.

[3] A. Farquhar, R. Fikes, and J. Rice. The Ontolingua Server: A Tool for Collaborative Ontology Construction. *IJHCS*, 46(6):707–727, 1997.

[4] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

[5] G. A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[6] P. Ziegler and K. R. Dittrich. User-Specific Semantic Integration of Heterogeneous Data: The SIRUP Approach. In *1st Int. IFIP Conference on Semantics of a Networked World (ICSNW 2004)*, pages 44–64, 2004.

[7] P. Ziegler, C. Kiefer, C. Sturm, K. R. Dittrich, and A. Bernstein. Detecting Similarities in Ontologies with the SOQA-SimPack Toolkit. In *10th Int. Conference on Extending Database Technology (EDBT 2006)*, pages 59–76, 2006.

[8] P. Ziegler, C. Sturm, and K. R. Dittrich. Unified Querying of Ontology Languages with the SIRUP Ontology Query API. In *Datenbanksysteme in Business, Technologie und Web (BTW 2005)*, pages 325–344, 2005.