ISENS: A Multi-ontology Query System for the Semantic Deep Web

Abir Qasem
Lehigh University
19 Memorial Drive West
Bethlehem, PA 18015, USA
abir.qasem@gmail.com

Dimitre A. Dimitrov
Tech-X Corporation
5621 Arapahoe Avenue, Suite A
Boulder, CO 80303, USA
dad@txcorp.com

Jeff Heflin
Lehigh University
19 Memorial Drive West
Bethlehem, PA 18015, USA
heflin@cse.lehigh.edu

Abstract

We present ISENS, a distributed, end-to-end, ontology-based information integration system. In response to a user's query, our system is capable of retrieving facts from data sources that are found in the surface Semantic Web as well as in the Semantic Deep Web. Furthermore, it retrieves facts from sources where the data is not directly described in terms of the query ontology. Instead, its ontology can be translated from the query ontology using mapping axioms. In our solution, we use the concept of source relevance to summarize the content of a data source. Our system can then use this information to select the needed sources to answer a given query. Source relevance is general enough that it can be used with both the surface Semantic Web and the Semantic Deep Web. In this paper, we show how we have incorporated three particular Deep Web data sources into our system to enable answering queries by composing information from the integrated sources.

1 Introduction

The Semantic Web provides an infrastructure that has the potential to transform the Web to a true global knowledge medium. Ontologies, expressed in a standard logic language with formal semantics, can be used in concert with web data in order to develop powerful query systems. The research community and the industry have made significant progress toward realizing this vision. For example, the Web Ontology Language (OWL) is now an international standard [7]. One key bottleneck to the success of the Semantic Web is data acquisition. The Deep Web, on the other hand, makes up about 77.3% of the current Web data [3]. This data is mostly stored in relational databases; it has well defined structure and some basic semantics defined by their schema. Therefore, acquiring and adapting the Deep Web data for the Semantic Web is a less complex task than acquiring and adapting semi-structured surface Web data like web pages and blogs. For example, nontrivial issues like natural language processing, data discrepancies and ontology learning are not that critical in acquiring Semantic Web data from the Deep Web. The Deep Web can be an ideal source for Semantic Web data.

Although there is a wide array of Deep Web data sources available, each Deep Web data source usually focuses on a particular area of a user's interest. Even the so called federated search engines that combine the results of multiple Deep Web data sources usually follow this same paradigm [9]. For example, a federated search engine may combine the results of three Deep Web data sources that provide reviews of "Fargo" the movie, but will not provide information about say the city "Fargo". However, in reality the information need of a user can span many data sources. For example, a user may want to know about some geographical information of a city like its elevation, latitude/longitude etc., but in addition she may be curious about the city's political information and may want to see some satellite images of the city. Her query therefore could potentially span a geographical database like Geonames[1], a Wikipedia style database like DBPedia[2] and maybe some individual blogs that have imagery of the city she is interested in. It is well known that web scale information integration has many challenges to overcome [2]. The Semantic Web may provide an integration mechanism that will allow a more comprehensive use of the Deep Web data beyond the data provider's anticipation.

In this paper, we present a system that demonstrates some of the synergies between the Semantic Web and the Deep Web that we have hypothesized above. Our ISENS system is a distributed, end-to-end, ontology-based information integration system which takes a SPARQL[3] query as input and retrieves facts from data sources that are found in the surface Semantic Web (files in OWL and Resource

[1] http://www.geonames.org/

[2] http://dbpedia.org/

[3] http://www.w3.org/TR/rdf-sparql-query/

Description Framework[4] (RDF) format) as well as the Semantic Deep Web. We say a data source is a Semantic Deep Web data source if it is either a Semantic Web Knowledge base or can be wrapped with a Semantic Web interface. This definition is consistent with the definition provided by Jung An *et al.* [1].

In our solution, we use the concept of a source relevance to summarize the content of a data source. Our system can then use this information to select the needed sources to answer a given query. Source relevance allows the content of a database to be declaratively described without regards to a specific access mechanism (e.g. query language, protocol etc.). This abstraction allows us to treat the surface Semantic Web and the Semantic Deep Web in the same manner. In our system, we ground these descriptions in RDF. We refer to these RDF statements that express a source's relevance to a query as REL statements. We postulate that these REL statements will either be handcrafted by the data providers or auto-generated using various data mining techniques [4].

In addition to addressing this format/access heterogeneity, our system addresses the semantic heterogeneity between data sources as follows. It uses an adapted information integration algorithm and mappings expressed in OWL to reformulate the query into the ontologies that the data sources use. We note that this reformulation is not the focus of this paper. The details of the reformulation and the information integration algorithm are presented in Qasem *et al.* [8].

The rest of the paper is presented as follows. In Section 2, we describe the overall architecture of our system and its key functionality. In Section 3, we provide specific details of the source relevance concept and how it is implemented in our system. In Section 4, we discuss the Deep Web data sources for our experiment and issues we have encountered in the data preparation. Finally, we conclude and provide the highlights of our future work.

2 ISENS System

In this section, we describe the overall architecture and the functionality of our system. As depicted in Fig. 1, ISENS consists of three main components: GUI, OBII, and distributed sources of meta data. We assume that each data source commits to one or more OWL domain ontologies. We use OWL axioms to describe a map between a pair of related ontologies. The choice of OWL to articulate the alignments make these maps shareable via the Web. Note that it is unlikely that we will have alignments between all pairs of ontologies, but it should be possible to compose an alignment from existing alignments. We also assume that each source has a set of REL statements, which will be discussed further in Section 3.

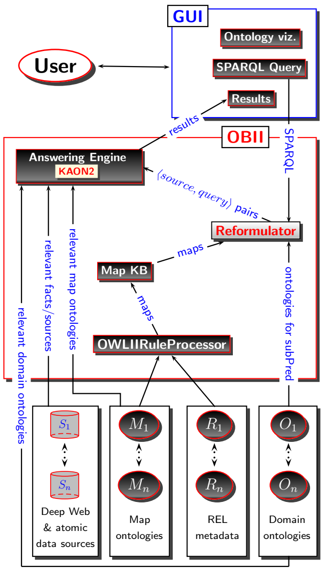

Figure 1. ISENS Architecture Diagram with arrows showing the flow of information when processing a query.

The primary component of our system is the Ontology Based Information Integrator (OBII). The OWLIIRuleProcessor loads the OWL maps and the REL statements. The Reformulator takes these inputs and implements our information integration algorithm. The AnsweringEngine is responsible for querying/loading the sources selected by the Reformulator into the KAON2[5] reasoner. In addition, the module loads the ontologies that are used in the reformulation and all the relevant maps. Then, it issues the original query to the reasoner and formats the retrieved answers.

Our Reformulator adapts the peer data management system (PDMS) reformulation algorithm [5] to the Semantic Web. The PDMS is a decentralized and extensible information integration architecture, in which any user can contribute new data, schema information, or even mappings between other peers' schemas. PDMS extends the standard Local-as-View (LAV) and Global-as-View (GAV) information integration approaches [6]. Since users can and will have queries in any ontology, we need to enable a mechanism that does not depend on a single mediated schema. Therefore, PDMS's "any schema" approach makes PDMS suitable for adaptation over data described with RDF/OWL ontologies. Qasem *et al.* [8] provide the details of our adaptation of this algorithm and defines an OWL-compatible mapping language to describe peer maps and sources.

Users interact with ISENS via its graphical user interface (GUI) component. The GUI allows loading of ontologies

[4] http://www.w3.org/RDF/

[5] http://kaon2.semanticweb.org/

available on the Web to visualize their class and property taxonomies (displayed as trees). Terms from the displayed (together with previously loaded and cached) ontologies can then be selected and used in constructing SPARQL queries in an interactive way. Once a SPARQL query is constructed, it can be submitted to the OBII component to find the relevant answers from the available data sources. The results found by OBII are returned back to the GUI and displayed.

3 Source Relevance

In Section 1, we introduced the notion of a REL statement. Due to the Web's size, it is infeasible to query all the web data sources for a given query. Therefore, we need some means of specifying meta data for sources. If we can determine which sources are potentially relevant simply from the meta data, then we can limit our queries to these sources. Having relevant information, however, does not mean that the source is capable of answering the query completely. It just says that the source may have some useful information on the query.

REL statements are formed using a subset of OWL and can be translated into LAV rules. Consider the statement REL(http://sourceURL, Electronics, CinemaDisplay $\sqcap \exists$ madeBy."DELL"). We omit the OWL syntax here for space consideration. It says that in a data source located at http://sourceURL there are some individuals of class "Electronics" that are cinema displays made by Dell. In this example, the meta:container will be the class expression that defines CinemaDisplay $\sqcap \exists$ madeBy."DELL", the meta:contained will be Electronics and meta:source will be http://sourceURL. The meta:RelStatement will encapsulate these three predicates. Please see Qasem *et al.*[8] for a detailed description of the REL statements.

We have also mentioned that the REL statements allowed us to focus on the content rather than the access mechanism. In our system we provide a translator that would translate the contained query atom into the target database's query language. We have currently implemented a translator to the SERQL query language. This is the query language for Sesame[6], which we use to store Deep Web data for our experiments. We plan to implement translators for SPARQL and other query languages.

4 Data Sources

Currently, there are only a few true Semantic Deep Web data sources that are accessed through SPARQL endpoints[7]. Most of these endpoints are not yet very reliable. However, as mentioned in Section 1, it is relatively easy to convert regular Deep Web data sources to Semantic Deep Web data sources. In our initial experiment we have done just that. Here, we briefly describe the three sources that we have incorporated into our system. In each case, after translating the data to OWL, we loaded into its own Sesame knowledge base.

DBPedia is an effort to convert all the Wikipedia data into a web enabled structured format. The DBPedia however, is not in OWL format and therefore we had to process the data so that its instances are typed according to OWL classes. For our initial experiment, we have used DBPedia's "infobox" dataset that was extracted from the English version of Wikipedia. The infobox data is well structured and its granularity enables fine-grained queries over the data set. Furthermore, we have only focused on infoboxes related to the cities of the world. We harvested city specific data for about 33,000 cities from this source. The properties of these city instances however are not used uniformly. For example, one city may have latitude/longitude information, whereas another may not. We decided to use the following set of properties that cover a large number of instances: population, population as of, postal code, image map, established date and airport code.

The Geonames web site provides information on various geographic entities. The database contains over eight million geographical names and consists of 6.5 million unique features of 2.2 million populated places. We collected city specific data of about 70,000 cities from this source. The data is fairly clean and we have converted it to OWL with simple scripts. The properties of cities that we have collected from this data set are latitude, longitude, elevation and country code.

The third data source we incorporated in the system consists of NASA provided Spaceborne Imaging Radar-C/X-Band Synthetic Aperture Radar (SIR-C/X-SAR) imagery and related semi-structured textual data[8] on cities. The NASA data are in HTML format. We decided to harvest and generate meta data in OWL on the NASA's "Cities" SIR-C/X-SAR data set in order to permit possible integration with the DBPedia and Geonames data sets. Thus, these three data sets enable us to develop a meaningful and interesting use case scenarios for ISENS with non-trivial ontology maps for information integration from the separate data sources.

We designed and developed an OWL ontology to describe the NASA city data. Then, to generate ontology individuals, we wrote code to crawl the NASA data, extract relevant information from the fetched html pages, and generate the individuals in OWL in the terms of the developed ontology. The code generates individuals about the cities for which imagery was provided, image data individuals that contain information about the dates the images were

[6] http://www.openrdf.org/

[7] http://esw.w3.org/topic/SparqlEndpoints

[8] http://www.jpl.nasa.gov/radar/sircxsar/cities.html

taken on, IDs, URLs to the specific images, city names, and descriptions of the images. We harvested city specific data for 33 cities from the NASA source.

An example of an interesting information integration use case query given these data sources consists of asking the system to find all cities for which there are image data available and their altitudes and their population density. Such a query can be expressed in a straightforward way in SPARQL. For testing the system, we have written this query in terms of the NASA ontology. The individuals of this ontology contain data about cities and imagery on them but there is no available information about the altitudes of these cities or population density. However, ISENS uses its ontology maps to deduce that the `hasAltitude` property in the NASA ontology is equivalent to the `elevation` property in the Geonames one and `hasPopulation` in the NASA ontology is equivalent to the `population` property in the DBPedia. The system then, using REL statements, finds relevant data sources on elevation data (from the Geonames individuals), and population data (from the DBPedia individuals) to answer this query and provide the elevation and population data of the cities for which it also found NASA imagery data.

5 Conclusion and Future Work

In this paper we present the ISENS ontology based information integration system that works seamlessly with both surface Semantic Web and Semantic Deep Web data sources. Given a user query, ISENS retrieves facts from multiple data sources, even from the sources where the data is not directly described in terms of the query ontology; and then integrates them to meet the user's information need. We have described how we have developed and incorporated three particular Deep Web data sources into our system to enable answering queries that provide complementary information from the integrated sources.

This work opens up some interesting avenues for further research. One of our short term goals is to investigate the robustness of our system in the case when a Deep Web data source is currently unavailable. Although we expect there will be natural redundancy of sources in terms of the information they provide, it is conceivable that in some cases we will not have any source available that can provide a certain piece of information. In that case the user query can only be partially answered. We want to perform a trade off analysis between providing partial results and having to postpone a query because one of its components is unavailable. A related area to look into is the optimization of the calls to the data sources. Depending on the query and the speed of the server, information about some components of the query will take longer to reach our system. We want to perform a similar analysis to come up with heuristics that will allow our system to handle the delay in a graceful and useful way.

A more long term research goal is to enhance the REL statements to accommodate web services. Web services expose their content through APIs. Unlike the Deep Web data sources, whose relevance can be modeled using our REL statements by specifying the expected result (i.e. output) of a query, web service API calls will also need inputs. We plan to investigate the changes required in our formal model to accommodate this enhancement.

6 Acknowledgment

We are grateful to the U.S. Department of Energy for funding this work under the DE-FG02-05ER84171 SBIR grant.

References

[1] Y. J. An, J. Geller, Y.-T. Wu, and S. A. Chun. Semantic deep web: automatic attribute extraction from the deep web data sources. In *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, pages 1667–1672, New York, NY, USA, 2007. ACM.

[2] N. Ashish and C. Knoblock. Semi-automatic wrapper generation for internet information sources. In *Proceedings of the Second IFCIS Conference on Cooperative Information Systems (CoopIS)*, Charleston, SC, 1997.

[3] K. C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang. Structured databases on the web: observations and implications. *SIGMOD Rec.*, 33(3):61–70, 2004.

[4] C. Halaschek, B. Aleman-Meza, I. Arpinar, and A. Sheth. Discovering and ranking semantic associations over a large rdf metabase. In *(Demonstration Paper)Proceedings of VLDB 2004*, 2004.

[5] A. Halevy, Z. Ives, D. Suciu, and I. Tatarinov. Schema mediation in peer data management systems. In *Proc. of ICDE*, 2003.

[6] M. Lenzerini. Data integration: a theoretical perspective. In *PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, New York, NY, USA, 2002. ACM.

[7] P. Patel-Schneider, P. Hayes, and I. Horrocks. OWL web ontology language semantics and abstract syntax. Proposed Recommendation, December 2003. http://www.w3.org/TR/2003/PR-owl-semantics-20031215/.

[8] A. Qasem, D. A. Dimitrov, and J. Heflin. Efficient selection and integration of data sources for answering semantic web queries. In *ICSC 08: Proceedings of the Second IEEE International Conference on Semantic Computing*. IEEE Computer Society Press, 2008.

[9] L. Si and J. Callan. Modeling search engine effectiveness for federated search. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 83–90, New York, NY, USA, 2005. ACM.