Efficient Knowledge Management by Extending the Semantic Web with Local Completeness Reasoning

Abir Qasem and Jeff Heflin
Dept. of Computer Science & Engineering
Lehigh University
19 Memorial Drive West
Bethlehem, PA 18015
{qasem , heflin}@cse.lehigh.edu

1 Introduction

In this post-industrial "information economy" more and more organizations are realizing that having actionable information gives them an invaluable competitive advantage. The term "intellectual asset" has been coined to reflect the significance of this type of information. The field of knowledge management (KM) provides tools, techniques and processes for the most effective use of an organization's intellectual assets [Davies 00]. But the advent of the Web and its subsequent ubiquity has fueled a rapid growth in information volume that has not slowed down. In fact, in addition to the more traditional web pages, we now have information from diverse sources like databases, sensors, web services and even intelligent agents. This diversity of data sources combined with the trends in division of labor in modern companies lead to a knowledge space that is highly distributed and ever changing. The traditional KM tools assumed a centralized knowledge repository and therefore are not suitable for this distributed knowledge medium [van Elst et. al. 03]. New KM tools are needed that integrate the knowledge sources dispersed across web into a coherent corpus of interrelated information.

The Semantic Web offers a more suitable platform for information integration then the traditional Web. Since the data has well-defined meaning, software, instead of humans can be used to harvest information and subsequent knowledge from a wide variety of sources. [Davies et. al. 03] postulate that it will significantly improve the acquisition, storage and retrieval of organizational knowledge. They propose an architecture for KM in the Semantic Web that addresses all aspects of KM lifecycle, namely acquisition, representation, maintenance and use.

In the use phase, efficiency of knowledge retrieval is of paramount importance. This is difficult to attain in the Web's distributed knowledge space because information from several diverse sources that have different capabilities and communication protocols need to be pulled together in a timely fashion for it to be useful to the organization. Also, due to the large quantity of information sources, we cannot afford to query all of them. However, if we have compact meta-level descriptions of each source, then we can determine which set of sources need to be accessed, resulting in more efficient queries. In addition to providing information on relevant sources, source descriptions may also be used to indicate when accessing a source would be redundant. Such a redundancy might

occur when two different sources have overlapping content. Reasoning about overlap is important for efficient KM.

In this work we use a formalism to characterize this overlap so that we can reason with it. Following work in the field of information integration [Friedman and Weld 97], our formalism is based on Local Closed World reasoning. We have augmented the W3C web ontology language (OWL) to allow us to express LCW statements. We have then described information sources using this augmented language and developed a prototype system that reasons with overlapping information and provides an integrated knowledge space to the user.

The rest of the paper is organized as follows: In Section 2 we provide pointers to related work, in Section 3 we describe our theoretical work and a simple prototype system that we have built, and in Section 4 we provide our conclusion and areas for future work.

2 Background

To be able to reason with overlap, we need to characterize and exploit source overlap. Addressing the source overlap problem is based on the idea of Local Closed World (LCW) information. LCW as proposed by [Golden et. al. 94] is a formalism for obtaining closed-world information on subsets of information that is known to be complete (LCW is described in more details in section 3). This formalism still allows other information to be treated as unknown. [Levy 96] extended this formalism to obtain complete answers from databases that have incomplete information.

All of this work however, assumes a priori knowledge of the local completeness information for each information source, which is an invalid assumption in the case of the Web. The Semantic Web on the other hand provides interesting possibilities for content providers to advertise the completeness of their sources. [Heflin and Munoz-Avila 02] have demonstrated this in a plan generation problem. They have shown how a planner can exploit LCW information encoded in SHOE [Heflin 98], another Web ontology language developed at University of Maryland, to complete a plan. Our paper builds on this work. We extend OWL to express source completeness and develop a system that reasons with that characterization and selects appropriate sources with respect to a query. Our system is based on the concept of "mediators" proposed by [Wiederhold 1992]; it is a system that is capable of integrating multiple sources in order to answer questions for another system.

3 Expressing Completeness on the Semantic Web

LCW information is used to specify the subsets of the information in a knowledge base that are known to be complete, while other information can still be treated as unknown. LCW information is given as meta-level sentences of the form LCW(ϕ), where ϕ is a first order logic sentence that contains one or more variables. If a sentence matches a substitution for ϕ then it is either already entailed by the knowledge base or it is false. In this sense, the sentence ϕ provides a scope for the relative completeness of the knowledge

base. Note, that this information is local in the sense that it is local to the knowledge base that it describes.

However, these first order logic (FOL) formulas cannot be directly adapted to the Semantic Web. OWL, the de facto standard for the Semantic Web is closer to Description Logic (DL) rather than FOL. To represent LCW using OWL one has to express the formulas in DL. Unlike FOL, which allows us to refer to an object, DL only has notation to express definitions and properties of classes of objects. Classes provide an abstraction mechanism for grouping objects with similar characteristics. OWL classes are described through "class descriptions". Hence we have to express LCW for a class description, which will mean that we have LCW over all the instances of that class.

We use two meta-level statements to characterize a source's ability to provide complete information with respect to a query. They express the relevance information and LCW information about the contents of each information source. They are represented by formulas of the form LCW (ϕ) and REL(ϕ). For a knowledge source i, $LCW_i(\phi)$ indicates that for all x, if i does not entail that x is of type ϕ, then x is in the complement of ϕ. Formally, we can say $\forall x\ i \not\models type(x, \phi) \Rightarrow x \in \phi'$. $REL_i(\phi)$ on the other hand indicates that there exists an instance o in i such that i is of type ϕ. Formally, $\exists o$ instance(o) $\wedge$ in (o,i) $\wedge$ type(i, ϕ).

We propose that an OWL document can use new properties lcw: isCompleteFor and lcw: isRelevantFor to state that it has complete or relevant information on some subset of information respectively. These properties are in a new namespace identified by the lcw prefix, and has rdf:Resource in its domain and owl:Class in its range. As such, it can be applied to any resource. The following examples show how to apply these properties to represent LCW. REL statements are expressed in a similar way.

We use the following to represent LCW (p (x, c)) on source s.

```
<rdf:Description rdf:about="s">
     <lcw:isCompleteFor>
         <owl:Restriction>
             <owl:onProperty rdf:resource="#p" />
             <owl:hasValue rdf:resource="#c" />
         </owl:Restriction>
        </lcw:isCompleteFor>
 </rdf:Description>
```

Here `isCompleteFor` is applied to an individual of a class that has at least one of its property values equal to the resource `c`.

It is somewhat difficult to represent complete information on an object's values for a specific property. So to represent LCW (p (c, x)), we create an anonymous property that is the inverse of p, and restrict the value of the inverse (essentially restricting the value of

the subject of p). To represent LCW (p (x,y)) we simply identify any individual with a property p.

We have built a prototype system, the Semantic Web Mediator, which identifies appropriate information sources with respect to a query. Its knowledge base contains two kinds of meta-information on the queries. It stores completeness information (i.e. which sources have all possible information about a query) and it stores relevance information (i.e. which sources have some information about a query). Using REL information, it can be determined which sources are relevant to a specific query. The LCW information makes it possible to prune redundant sources from this set. The knowledge base itself is initialized from OWL files that contain explicit description of sources' ability to provide completeness information.

4 Conclusions and outlook

In our work we have adapted LCW, a formalism commonly used to find relevant answers from an incomplete database, to characterize redundant information on the Semantic Web. We postulate that this representation will increase efficiency of KM on the Semantic Web. We have built a proof of concept system to explore the feasibility of this concept.

We are now in the process of building a more complete system. We plan for the following in recent future:

a) Provide support for complex queries and data sources that commit to heterogeneous ontologies
b) Allow for dynamic update of the system's knowledge base

5 References

[Davies et. al. 03] Davies, J.; Fensel, D.; Van Hermelen, F. 2003. Towards The Semantic Web: Ontology Driven Knowledge Management. John Wiley & Sons, NJ.

[Davies 00] Davies, J. 2000. *Supporting Virtual Communities of Practice*, In Roy, R. (ed.), Industrial Knowledge Management, Springer Verlag.

[Friedman and Weld 97] Friedman, M. and Weld, D. 1997. Efficiently Executing Information Gathering Plans. In *Proc. of IJCAI-97*.

[Golden et. al. 94] Golden, K.; Etzioni O.; and Weld, D. 1994. Omnipresence Without Omniscience: Efficient Sensor Managment for Planning. *In proc. of AAAI-94*.

[Heflin 98] Heflin, J.; Hendler J.; and Luke S. Reading Between the Lines: Using SHOE to Discover Implicit Knowledge from the Web. 1998. In *AI and Information Integration. Papers from the1998 Workshop. WS-98-14.* AAAI Press, Menlo Park, CA, 1998. pp. 51-57.

[Heflin and Munoz-Avila 02] Heflin, J. and Munoz-Avila, H. 2003. LCW-Based Agent Planning for the Semantic Web. In *Ontologies and the Semantic Web. Papers from the 2002 AAAI Workshop WS-02-11*. AAAI Press, Menlo Park, CA, 2002. pp. 63-70.

[Levy 96] Levy, A. 1996. Obtaining Complete Answers from Incomplete Databases. In *Proceedings of the 22'nd VLDB Conference.*

[OWL 04] *OWL Web Ontology Language Guide*, retrieved March 15th, 2004 from http:// http://www.w3.org/TR/owl-guide/

[van Elst et. al. 03]. van Elst, L., Dignum V., Abecker, A 2003. Agent Mediated Knowledge Management, Springer Verlag.

[Wiederhold 92] Wiederhold, G. 1992. Mediators in the Architecture of Future Information Systems. *IEEE Compute*r.