SWWS 2001 - Position PaperÝ

Supporting Knowledge Discovery on the Semantic Web by Exploiting the Semantics of Complex Relationships

K. Anyanwu, A. Sheth,Ý [á]

[á] LSDIS Lab {anyanwu,amit@cs.uga.edu.

Department of Computer Science, University of Georgia, Athens, GA 30602, USA,

Summary

Current research efforts into the role of ontologies in the Semantic Web have focused mainly on semantic modeling, querying, information exchange and integration. Correspondingly, most of the specification languages including DAML/OIL, as well as other XML/RDF based languages, e.g. SHOE, provide support for representing basic semantic relationships between ontology concepts, typically in a single domain. The reasoning support provided for systems is limited to that supported by description logic, i.e., subsumption, which may be used to derive relationships of a hierarchical nature like is-a relationships, and sometimes instance-of relationships.

Focus of our research is on knowledge discovery and complex information requests that may involve correlating information from disparate domains. The entities in the different domains may have complex relationships e.g. causal relationships whose semantics cannot be modeled using current ontology representation languages.Ý In this position paper, we motivate the need for support of complex relationships with a scenario. We outline how our approach for supporting inter-domain relationships together with a rich query mechanism can be used to support knowledge discovery.

Background

The modeling primitives available in present-day ontology languages allow for expressing mainly hierarchical relationships (inheritance) and relational properties like transitivity, symmetry etc. This affords a reasonable level of reasoning capability, but is somewhat limited in the nature of questions that can be answered by systems. For example, questions of the is-a, part-of, instance-of, nature may be readily answered by such systems.Ý In contrast, a system like InfoQuilt that supports semantic information correlation [2] with more complex relationships both within and across domains can be used to answer more exploratory questions. For example, one question that has recently become of interest to geography researchers is ìDo Nuclear Tests cause Earthquakes?î. Considering the wealth of information on the web, it should be possible for a Semantic Web solution to correlate and analyze information from federated (heterogeneous and autonomous) information sources containing information on nuclear tests and those containing information on earthquakes, and come up with a preliminary answer to that question. The capability that is needed here, and is provided by InfoQuilt, goes beyond the traditional approach of providing integrated views of multiple data sources. It involves the ability to express and represent complex inter-domain relationships that can be exploited by the system to perform useful correlation amongst the different domain sources, as well as a rich information request mechanism that allows users to express more meaningfully their information need.

In the next section, we will outline how the unique features of the InfoQuilt system can be used to support complex relationships and knowledge discovery.

Complex Relationship Support In InfoQuilt

The aforementioned example explores the hypothetical causal relationship between Nuclear Tests and Earthquakes. In the InfoQuilt system, the user can explicitly represent detailed semantics of this relationship, using attributes/properties from the Nuclear Tests and Earthquake ontologies, and a library of operators and functions. The notation that is used below is a concise representation of the domain ontologies, containing domain attributes, attribute properties, and rules.

NuclearTestÝ (Ý testSite, explosiveYield, waveMagnitude, testType, eventDate, conductedBy, latitude, longitude,ÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝ
ÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝ waveMagnitude > 0,Ý waveMagnitude < 10,Ý testSite -> latitude longitudeÝÝ);
Earthquake (ÝÝ eventDate, description, region, magnitude, latitude, longitude, numberOfDeaths, damagePhoto, ÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝÝ magnitude > 0ÝÝ);
The following is a representation of the causal relationship.

NuclearTestCausesEarthquake :Ý <= dateDifference(NuclearTest.eventDate, Earthquake.eventDate) < 30
ÝÝÝÝÝÝÝÝÝÝÝ
ÝÝÝAND
distance(NuclearTest.latitude, NuclearTest.longitude, Earthquake,latitude, Earthquake.longitude) < 100

This relationship can be verbalized as follows: A nuclear test may be said to have caused an earthquake, if the earthquake occurred within thirty days and 100 miles of the test explosion.

Additionally, InfoQuilt provides a construct called an Iscape or Information Scape, which is used represent a userís information request. An Iscape is semantically richer than traditional keyword based or attribute based (e.g., SQL) query mechanisms because it contains information from the domain ontologies as well as any specified relationships between the domains to process and evaluate an information request. Consider the following text version of an Iscape.

ìFind all nuclear tests conducted by USSR or US after 1970 and find any information about any earthquakes that could have potentially occurred due to these tests.î

To answer this question, the query planner extracts information from only relevant nuclear test information sources, as well as relevant information sources on earthquakes. Then using information about the *NuclearTestCausesEarthquake* relationship, will compare dates and locations of nuclear tests and earthquakes and eliminate those that do not meet the relationshipís constraint [1,3].

Knowledge Discovery Using InfoQuilt

The framework provided by InfoQuilt can be used to support knowledge discovery either by formulating complex information requests. Alternatively a user may pose a hypothesis involving complex relationships between data from heterogeneous and autonomous Web-accessible information sources. Corresponding results can help either justify or falsify their hypothesis and guide further requests.Ý For example, to explore the aforementioned relationship, we can try the following sequence of Iscapes

ì*Find when the earliest recorded nuclear test was conducted.*î

We find from the results that nuclear testing began in 1950. So we use a few more Iscapes whose results show that there is a sudden increase in the number of earthquakes since 1950, and that in the period 1900-1949, the average rate of earthquakes was 68 per year and that for 1950-present was 127 per year, that is, it almost doubled. Next, we try to analyze the same data grouping the earthquakes by their magnitudes.

ì*For each group of earthquakes with magnitudes in the ranges 5.8-6, 6-7, 7-8, 8-9, and magnitudes higher than 9 on the Richter scale per year starting from year 1900, find the average number of earthquakes.*î

The results show that the average number of earthquakes with magnitude greater than 7 on the Richter scale have remained practically constant over the century (about 19).

We can therefore deduce that the earthquakes caused by nuclear tests usually are of magnitudes less than 7 on the Richter scale. We can then try to explore the data at a finer level of granularity by trying to look for specific instances of earthquakes that occurred within a certain period of time after a nuclear test was conducted in a near by region.

Conclusion

InfoQuilt provides support for representing and utilizing (1) domain knowledge including concepts, relationships, domain rules and data dependencies, (2) complex inter-ontology relationships, (3) a semantically rich information request mechanism, (4) modeling of information resources which captures the nature of content present, and (5) a library of operators and functions (user-defined) that are useful in defining semantic relationships as well as resolving syntactic heterogeneities.

We believe that taken together, these components support *deeper* semantics and provide a framework for supporting (defining, sharing, executing) semantic information correlations and complex semantic relationships between data managed by Web-accessible heterogeneous and autonomous information, such as InfoQuiltís Iscapes, should be investigated to future enrich the rapidly evolving Semantic Web.

References (also see: http://lsdis.cs.uga.edu/proj/iq/iq_pub.html)

1. S. Patel, and A. Sheth, "Planning And Optimizing Semantic Information Requests Using Domain Modeling And Resource Characteristics". to appear in 6th Intl Conf on Cooperative Information Systems (CoopIS 2001), Italy. Spetember 2001.
2. K. Shah and A. Sheth, "Logical Information Modeling of Web-accessible Heterogeneous Digital Assets", Proc. of the Advances in Digital Libraries," (ADL'98), Santa Barbara, CA, May 28-30, 1998.

3. S. Thacker, A. Sheth, and S. Patel, Complex Relationships for the Semantic Web, to appear in Creating the Semantic Web, D. Fensel, J. Hendler, H. Liebermann, and W. Wahlster (eds.) MIT Press, 2001.