Consensus Ontologies for Semantic Reconciliation

Larry M. Stephens and Michael N. Huhns
Department of Computer Science and Engineering
University of South Carolina, Columbia, SC, 29208, USA

Position Statement

Small, independently-developed ontologies can be related *without* the necessity of constructing a global ontology beforehand. We assume that Web pages for specific domains will be annotated with ontological information. Given these annotations, we have developed a methodology that merges individual ontologies into what we call a *consensus ontology*, which has the appearance of a global ontology for a particular query. The consensus ontology may be cached for later use. This approach allows consideration of additional information sources incrementally.

Information Retrieval

Information searches can involve data and documents both internal and external to an organization. The research reported at this workshop targets the following basic problem: a search will typically uncover a large number of independently developed information sources—some relevant and some irrelevant; the sources might be ranked, but they are otherwise unorganized, and there are too many for a user to investigate manually. The problem is familiar and many solutions have been proposed, ranging from requiring the user to be more precise in specifying search criteria, to constructing more intelligent search engines, or to requiring sources to be more precise in describing their contents. A common theme for all of the approaches is the creation, use, and manipulation of ontologies for describing both requirements and sources.

Unfortunately, ontologies are not a panacea unless everyone adheres to the same one, and no one has yet constructed an ontology that is comprehensive enough. Moreover, even if one did exist, it probably would not be adhered to, considering the dynamic and eclectic nature of the Web and other information sources.

There are three approaches for relating information from large numbers of independently managed sites: (1) all sites will use the same terminology with agreed-upon semantics (improbable), (2) each site will use its own terminology, but provide translations to a global ontology (difficult, and thus unlikely), and (3) each site will have a small, local ontology that will be related to those from other sites—our position.

The experimental methodology we developed relies on the ontological annotation of Web pages—a representation consistent with visions for the Semantic Web. However, a pre-existing global ontology is not required. The domains of the sites must be similar—else there would be no interesting relationships among them—but they will undoubtedly have dissimilar ontologies, because they will have been annotated independently.

Experimental Methodology

We assigned graduate students in computer science the task of constructing small ontologies for three domains. The ontologies are written in DAML/OWL and contain at least 8 classes organized with at least 4 levels of subclasses.

We merge the individual files for a particular domain one-at-a-time into a resultant merged file. Node merging is based on syntactic and semantic information. The syntactic information is derived from the names of the nodes, for which we employ various string-matching techniques including detection of plural endings. The semantic information includes the meaning of the subclass link in the ontologies, prefixes that indicate antonyms, and evolving sets of synonyms for matching nodes. The synsets, which are used to track the progress of merging and to monitor correctness, are seeded from WordNet.

For each node in the resultant file, we maintain a *reinforcement* value, which indicates how many times the node is matched as ontologies are merged. We also maintain reinforcement values for class-subclass links. Next, we construct a *consensus ontology* by eliminating weakly reinforced nodes and links. In filtering the merged file, we sort the subclass *links* by their reinforce-

ment values and find that, for the most part, the strongly reinforced nodes are associated with strongly reinforced links. This finding, while not surprising, makes constructing a consensus ontology more efficient.

The software for merging ontologies can be found at http://www.cse.sc.edu/research/cit/projects/DAML.html. Sample ontologies are also available on that page.

Challenges

Our work focuses only on the *class-subclass* relationship among concepts. Other relationships such as *partOf* offer semantics that can be exploited in generating and restricting a consensus view.

Biographies

Larry M. Stephens received the B.S. degree in electrical engineering from the University of South Carolina in 1968 and received the M.S. and Ph.D. degrees in electrical engineering from the Johns Hopkins University in 1974 and 1977, respectively. He is currently Professor of Computer Science and Engineering at the University of South Carolina and a member of the Center for Information Technology. Prior to his academic career, he served for four years as a U.S. Naval officer assigned to the Naval Reactors Program, Washington, DC.

From 1988 to 1989 he was on leave as a consultant to the Microelectronics and Computer Technology Corporation (MCC), Austin, Texas. At MCC he conducted research on distributed knowledge-based systems, reasoning architectures for synthesis tasks, and plausible inferencing based on the properties of semantic relations. At MCC, he did extensive work with the Cyc common-sense reasoning project. His current research interests include the fundamentals of knowledge representation, design of ontologies, and information retrieval from heterogeneous information sources.

Michael N. Huhns received the B.S.E.E. degree in 1969 from the University of Michigan, Ann Arbor, and the M.S. and Ph.D. degrees in electrical engineering in 1971 and 1975, respectively, from the University of Southern California, Los Angeles. He is a Professor in the Department of Computer Science and Engineering at the University of South Carolina, where he also directs the Center for Information Technology. Prior to this, he conducted research on the Argo, Antares, RAD, Carnot, and InfoSleuth projects at the Microelectronics and Computer Technology Corporation as a Senior Member of the Research Division. He was also an adjunct professor in computer sciences at the University of Texas. Before joining MCC, he was an associate professor at the University of South Carolina, a research assistant in image processing at the University of Southern California, and a radar systems engineer at Hughes Aircraft Company.

Dr. Huhns is currently teaching courses in information technology and conducting research in cooperative information systems, ontologies, cooperative information systems, software agent systems, and bioinformatics.