Information Retrieval on the Semantic Web*

Urvi Shah and **Tim Finin** and **Yun Peng**
University of Maryland
Baltimore County

James Mayfield
Johns Hopkins University
Applied Physics Laboratory

1 Introduction

We envision the future web as pages containing both text and semantic markup. Current information retrieval techniques are based on keyword searches and hence cannot give precise answers to precise questions. Knowledge representation languages like DAML+OIL that support logic inferences can help us achieve more flexible and precise information retrieval.

We describe an approach for information retrieval over documents that consist of both free text and semantically enriched markup statements in DAML+OIL. These statements provide both structured and semi-structured information about the documents and their content. Our approach allows inferencing to be done over this information at several points: when a document is indexed, when a query is processed and when query results are evaluated. To validate our approach we have implemented a working prototype which is based on a version of the HAIRCUT information retrieval system and uses a semantic web inferencing system implemented using DAMLJessKB.

2 Design and Implementation of HOWLIR

HOWLIR is intended to provide a framework, which is able to extract and exploit the semantic information from semantically marked documents, perform sophisticated reasoning and filter results for better precision. *Ontologies* support information retrieval based on the actual content of a page. HOWLIR defines ontologies encoded in DAML+OIL allowing users to specify their interests in different events and retrieve relevant information. The *Event Ontology* is built following the concept of "Natural Kinds OF" from the field of philosophy. We first identify the natural kinds in the phenomena under study, "EVENTS", and then figure out what their most important characteristics are.

We take advantage of the AeroText™ system for text extraction of key phrases and elements from event announcements which are currently in free text. The extracted phrases and elements with reference to the Event Ontology play a vital role in identifying type of events and adding semantic markup. We have built DAML generation components that translate the extraction results into a corresponding RDF triple model that utilizes the DAML+OIL syntax.

*This research was supported in part by DARPA contract F30602-97-1-0215. Extended version of this paper is available at http://daml.umbc.edu/papers/howlir2002.pdf

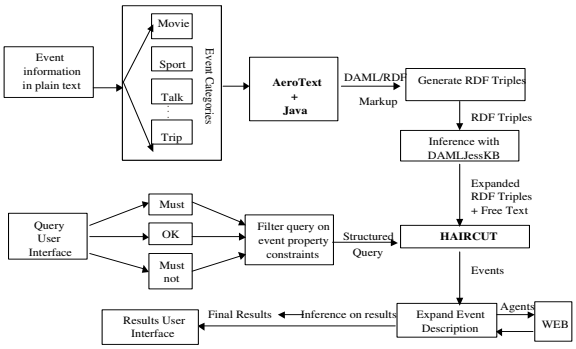

Figure 1: HOWLIR Framework

HOWLIR uses the metadata information added during the text extraction process to infer additional semantic relations that decide the scope of the search and to provide more relevant responses. HOWLIR bases its reasoning functionality on the use of DAMLJessKB. We enhance the existing inferential capabilities of DAMLJessKB and supplement it by applying domain specific rules for reasoning over instances and concepts of the Event ontology and filtering out facts that are of relevance to our system.

The Hybrid Information Retrieval mechanism is based on the use of JHU/APL's HAIRCUT System. Traditional text retrieval characterizes documents by the indexing terms they contain. We reduce document markup to RDF triples, and treat each distinct triple as an indexing term. The addition of semantic markup to Web documents makes it possible to perform inference over document content. Taking advantage of the HAIRCUT feature which allows the user to specify which terms in the query MUST, MUSTNOT and MAYBE considered, each query is expressed as a document consisting of triples and free text. This gives the user flexibility in querying, at the same time increases precision.

3 Conclusion

The HOWLIR framework for information retrieval over the Semantic Web utilizes a set of ontologies and a hybrid information retrieval mechanism. HOWLIR can be used to answer queries about explicit and implicit knowledge specified by the ontology thus provide a query answering facility that performs deductive retrieval from knowledge represented in DAML+OIL.