SAS Institute Inc. SAS Campus Drive Cary, NC 27513
Tel: 919-677-8000 Fax: 919-677-4444 www.sas.com

7th International Protégé Conference: Presentation Abstract
Eloise Currie, Application Developer
Mary Parmelee, Metadata Specialist

Toward a Knowledge – Based Solution for Information Discovery in Complex and Dynamic Domains

Introduction

In recent years, the rapid expansion of online information has created a new media venue that is simultaneously one of the most promising and challenging in history. As the volume and complexity of online information grows, retrieving relevant information in a timely fashion has become a major challenge. Moreover, rapidly changing domain terminology exacerbates the semantic mismatch between natural language and computer systems.

Both search queries and online resource information are communicated via natural language, which is defined by two main properties, syntax and semantics. Syntax gives language structure and order, while semantics give the context-sensitive meaning of the terminology within language. Yet traditional search engines match only on syntax, returning relevant information that must be manually filtered from tens or even hundreds of irrelevant results that are literally taken out of context.

Using emerging semantic technologies, SAS has recently developed a semantic information retrieval system for SAS® software online documentation. This knowledge-based system enhances the user's search experience in three ways: by selectively filtering out irrelevant information from the results set, by employing fuzzy matching techniques to provide support for common misspellings and synonyms, and by adding a contextual browsing mechanism that enables the user to drill down to a desired level of granularity.

SAS Institute Inc. SAS Campus Drive Cary, NC 27513
Tel: 919-677-8000 Fax: 919-677-4444 www.sas.com

This presentation describes our system development progress to date, the technologies that we implemented, the major obstacles that we encountered and how we overcame them. We demonstrate the current Knowledge Base system, focusing on how we leverage the semantics captured in ontologies to enhance information retrieval. Finally we share our vision for an integrated solution where syntax and semantics work in concert to provide a complete content development and knowledge delivery solution.

System Development

The system development process consists of three main areas: Knowledge Base development, Knowledge Base deployment and query, and the Knowledge Delivery system.

Knowledge Base development defines an intelligence layer to capture the meaning of resource content and provide a framework for information delivery. It is a seven step process that uses Protégé as a central technology for semi-automated ontology generation and Knowledge Base instance population. This process incorporates five custom Protégé plugins as well as the SAS Text Miner product for hierarchical resource clustering.

The Knowledge Base deployment and query process applies the intelligence layer to resolve semantic difference using category hierarchies to provide context and advanced search techniques. It has two main stages: first we use the Protégé API to generate a persistent MySQL Knowledge Base; then we use the Algernon inference engine for Knowledge Base Object query and navigation. Algernon matches Knowledge Base Objects (classes and instances) to user-defined search criteria and implements fuzzy matching techniques that handle misspellings, synonymous phrases and alternative word forms. Algernon also blends multiple hierarchies to produce browsable taxonomy and bread crumb trail views of Knowledge Base Objects in the Knowledge Delivery system.

SAS Institute Inc. SAS Campus Drive Cary, NC 27513
Tel: 919-677-8000 Fax: 919-677-4444 www.sas.com

The Knowledge Delivery system is built as a J2EE web application using the Struts Tiles framework which provides a model-view-controller framework for web application. It delivers information in context using browsable categories, categorized search results, hover text descriptions, category bread crumb trails, contextual category and full text search. The Ontology Browser implements the taxonomic tree view of Knowledge Base Objects. In addition to knowledge object and fuzzy matching techniques, the Knowledge Delivery system employs search filters to limit the search field by user- specified preferences and search limiters. The open-source Java library Lucene, is used to expand search functionality by providing contextual full text search capability that broadens the keyword search within the context of a given taxonomy node.

Vision for an Integrated Solution

Future system development will focus on an integrated solution where syntax and semantics work in concert to provide a complete content development and knowledge delivery solution. Content development plans include an XML-based structured architecture that supports a modular writing process for SAS content development. Modular writing divides content into semantically distinct modules for the purpose of optimizing reuse and reducing authoring redundancy. The shift to modular writing will simplify Knowledge Base development by enabling us to model semantically distinct content modules instead of complex resource objects. The system will dynamically assemble modules into complex resource objects that are relevant to the context of the current query. In order to support dynamic assembly of complex resource objects and enhance inference capability, we will shift to a more formal semantic representation using the OWL Web Ontology Language.