Fall 2005 ORCA Mentoring Grant Application

Semi-Automatic Generation of Mini-Ontologies from Canonicalized Tables

First applicant name: **Chris Hathaway**
First applicant email and RouteY login: **chris_hathaway@byu.edu ckh27**

Mentor name and department: **Dr. David Embley, Computer Science**

Importance of Project

My project is one of three integral parts of the TANGO initiative (Table ANalysis for Generating Ontologies), a collaboration between Brigham Young University and Rensselaer Polytechnic Institute. The goal of TANGO is to offer an environment for the semi-automatic production of ontologies from structured data, which will provide useful resources for the implementation of the Semantic Web. My research will connect the extraction of the tables from the web to the combining of the mini-ontologies I create into a larger domain ontology.

Main Proposal Body

With the explosion of online data and resources in recent years, finding relevant information given a user's query is becoming increasingly difficult. The available information is certainly only as useful as it is obtainable. Current search techniques base their results on keywords and word locations, which are often less helpful because of ambiguities in human languages. To get anywhere beyond current capabilities, we must endow the computer with more than data; the machine must be able to understand semantics and utilize them in a way that imitates intelligence. These goals are realized in the vision of the Semantic Web, but many intermediate tools must be created and many problems overcome in the progress toward a more usable and available Internet.

An important step along the way is the creation and maintenance of domain ontologies. An ontology is a machine-understandable representation of a human conceptualization [2]. For example, we know what a car is, and we know that cars have wheels, cost money, and need repairs. When this knowledge can be expressed in a more precise form for the machine, with its roots in predicate logic, it will become an ontology, readable by Semantic Web agents. These agents then make inferences and connections among several ontologies to carry out a given task.

While these ontologies are helpful in making information understandable to machines, their biggest setback lies in creation and upkeep. Currently, the best methods are the slowest, where experts in a field create the ontology according to their own knowledge. And because ontologies generated by individuals are often biased according to the opinions of the creator, reliable ontologies will only come through agreement among several experts. Concepts also change their properties, actions, and requirements over time, so constant maintenance is required to assure quality of the represented knowledge. These processes consume more time than is available, and lag far behind the intensifying amounts of resources appearing on the Web.

This problem necessitates the intervention of computing technology to make the methods automatic, or at least semi-automatic, with occasional help from an informed user. Many investigators have already recognized this, and have attempted to create methods of generation from current web data. The consistent theme in the troubles researchers have faced was that they were attempting to extract ontologies from free-form, unstructured text, which is full of the uncertainties of human language.

A more effective alternative, or perhaps simply an intermediate step, is to extract ontologies from the *structured* data on the web, such as charts or tables. When web authors create a table, they are forced to organize the information, which is often done in a predictable manner. A table need not look like the stereotypical grid with rows and columns of information. It can span over several web pages and be in different formats, such as a bulleted list. These tables, in all forms, will be taken and transformed into a standardized format that corresponds to a determined, canonicalized set of information similar to a relational database, with attribute-value pairs.

The project will then take the standard tables and using rules, inferences, assumptions, and predictions, extract the most probable set of objects, relationships and constraints for the given table, creating a mini-ontology, which comprehensively describes the data from the original source. I will further research several options in making this process as automated as possible while still maintaining a high level of accuracy in the resulting ontologies. In so doing, I will show that it is not only possible to extract ontological data from tables, but that the process can be automated to the point that the user needs little or no expertise in the table's domain. Although not part of my project, a later process will integrate the mini-ontologies into a large evolving domain ontology.

Anticipated Academic Outcome

The most tangible product of this research will be the tool that I plan to create to facilitate the described operations. But I plan to also produce a final paper in the form of a thesis to describe the specific methods used in the process and the analysis of the efforts to make the system more automated. I will use this thesis as part of the requirements to graduate with Honors, and the insights expressed will help future work on the TANGO project as a whole.

Qualifications

To prepare for research in the generation of ontologies as a step in enhancing the current Web, I have taken several classes in related Computer Science fields, such as Databases, Machine Learning, Distributed Systems, and most importantly, a graduate-level class on the use of Description Logics and Ontologies in the Semantic Web. It was in this last class that I did preliminary research for this project, including a paper that took the form of a thesis proposal. I have also completed a review of the various literature associated with this topic in order to better understand my objectives.

Dr. David W. Embley is leading the efforts of the TANGO project, of which my research will be a part, so he is the best choice for advisement as I perform the described investigation. The TANGO project was proposed by Dr. Embley, among others, and has received a three year grant from the National Science Foundation (Grant #0414644), used to support the graduate students working on the TANGO system. Dr. Embley has been involved in several publications relating to the Semantic Web and information extraction, including some specifically targeted at the generation of ontologies through the use of tables. His almost 30 years of teaching experience, including the overseeing of dozens of graduate theses and dissertations, should be very helpful to my research experience.

Project Timetable

Decide functionality	October 15, 2005
Create design/specification document for tool	October 29, 2005
Build framework (attached to the Ontology Editor)	November 26, 2005
Program all manual features	December 17, 2005
Phase 1 of automated features; analyze effectiveness	January 28, 2006
Phase 2 of automated features; analyze effectiveness	February 25, 2006
Write first (edited) draft of complete thesis	March 15, 2006
Have polished draft ready for review	April 1, 2006
Submit final report	June 1, 2006

Fit With BYU's Mission [optional]

Not Applicable

Sources

[1] Yuri A. Tijerino, David W. Embley, Deryle W. Lonsdale and Yihong Ding, "Towards Ontology Generation from Tables," *World Wide Web Journal: Internet and Web Applications*, (in press).

[2] M. Gruninger and J. Lee. "Ontology applications and design," *Communications of the ACM*, 45(2):39-41, February 2002.

[3] Y. Ding and S. Foo, "Ontology research and development. Part 1 - A review of ontology generation," *Journal of Information Science*, 28(2), 123-126, 2002.

[4] H. Mannila and K.-J. Räihä. "Algorithms for Inferring Functional Dependencies," *Data & Knowledge Engineering*, 12(1):83-99, February 1994.

[5] Biskup, J. and D. Embley: "Extracting Information from Heterogeneous Information Sources Using Ontologically Specified Target Views," *Information Systems*, 28(3): 169-212, 2003.