Can AI Techniques be Applied to Forest Science Data Integration Problems?

Steven B. Dolins[1], Richard Condit[2], Hua-Ching Su[3], Suzanne Lao[2]

[1]Bradley University
Department of Computer Science and Information Systems
Peoria, IL. 61625
sdolins@bradley.edu

[2]Smithsonian Tropical Research Institute, Panama City, Panama
[3]Independent Consultant, Milpitas, CA.

Abstract

A workshop meeting was held December 17-18, 2006 at the Center for Tropical Forest Science to discuss database technology and forest science. Approximately twenty botanists participated in the meeting, representing twelve research organizations that collect data from seven plots or surveys from around the world. At the meeting we discussed technical issues, such as how to share existing schemas, how to integrate and share data, and how to represent botanical taxonomies. A number of the data requirements and possible solutions that were discussed at the meeting are briefly described in this paper. The objective of this paper is to explore whether current semantic web tools can provide practical solutions for integrating, representing, and sharing heterogeneous data sources.

Introduction

The Center for Tropical Forest Science (CTFS), part of the Smithsonian Tropical Research Institute (STRI) in Panama, monitors and studies a global network of rain forest census plots, which include 6,000 tree species. Botanists at CTFS study ecological systems by collecting, storing, and analyzing large amounts of measurement data from these forests, specifically trunk diameter, location, and species for over five million individual trees. CTFS uses these data to compute growth, mortality, and recruitment rates for tropical tree species across different habitats and thus to address larger scientific questions about the impact of climate and atmosphere on the forests.

A workshop meeting was held December 17-18, 2006 at CTFS to discuss database technology. Approximately twenty botanists participated in the meeting, representing twelve research organizations that collect data from seven plots or surveys from around the world. Some of the plot or inventory databases built by these participants include SALVIAS, RAINFOR, Vegbank, and the US Forestry Inventory database (CTFS), (SALVIAS), (RAINFOR), (Vegbank), (US Forestry Inventory). The SALVIAS system is notable because it also stores taxonomic data and provides tools for taxonomic editing.

At this meeting a number of technical issues were discussed, such as how to share existing schemas, how to integrate and share data, and how to represent botanical taxonomies.

This paper will briefly describe these technical issues, offer conventional technical approaches, and list possible applications of AI. The objective is to create discussion points about requirements for scientific applications and the potential for AI applications.

Database Issues

At the December 2006 workshop, several botanists discussed the databases they designed and built (CTFS), (SALVIAS), (RAINFOR), (Vegbank), (US Forestry Inventory). Essentially, each database stored the same type of tree data, i.e., diameter, location, and species information. Although each database schema was similar, each one had its own idiosyncrasies. For example, some database schemas represent data about the number and diameter of each branch for every tree (a branch being anything larger than 1 cm diameter), but some botanists work in forests where the trees do not have any branches so this information is not captured in their schemas. Other differences include the number and types of attributes stored per tree and measurement.

Because the data requirements had major differences, it was unclear how to compare competing logical data models or how to answer the following question: "*Why create another model when several already exist?*"

Two possible solutions were discussed at the meeting: 1) adopt one, canonical model, possibly taking the best features from the different models or 2) keep different schemas but create one, canonical external view of the data. That is, define an external view which unifies the disparate logical models. The base tables can be plot specific with data idiosyncrasies for each plot. The external view can be made "canonical".

There are a number of advantages for having one, canonical schema: 1) different researchers could standardize their techniques for data collection and entry, 2) scientific studies could become more uniform, and 3) data can be easily shared. The disadvantages or problems with creating one canonical model are overwhelming: 1) getting scientists to agree on one formal representation is unrealistic, and 2) getting scientists to change their current schemas could result in a significant effort.

Another related technical problem is how the scientists can share data. That is, how can scientists obtain data and get it in a useable form? One solution to help scientists share data is to build a multidimensional model, i.e., design and build a data warehouse. A data warehouse is an integrated, static database only used for analysis purposes (Kimball 1996). Rather than store data in a network of disparate databases or try to fit all data into one schema, design an integrated, data warehouse for tree data. One can envision integrating data from various forests in a single database and also integrating other sources of data such as climate and soil information. In data warehouse design there are two types of tables, dimension and fact tables. In this domain, trees, geography, and censuses or time are possible dimensions and measurements are facts.

The advantages and disadvantages of a data warehouse are both significant. Having an integrated database would create a centralized data environment. This environment would make data easily available and facilitate research projects where comparisons of data are required. The problems may outweigh the benefits. There are numerous rain forest sites that are capturing data so scalability may be an issue. Hardware and performance issues would be difficult and there is no one central site that has the resources to undertake this effort. And although some data are accessible for sharing, there are also issues regarding researchers from different countries sharing their proprietary data.

Another approach that scientists can take to share data is to develop a software application programming interface (API); i.e., the API specifies an interface for extracting data. The API can extract data from "view" tables. As previously described, view tables are external to the logical design and these tables can be designed to support different types of users. Using an API does not preclude building a data warehouse for integration.

Developing an API has both advantages and disadvantages. An API is uniform and can hide details and complexities of the base tables in the various database designs. On the other hand, developing an API will require a software development effort as well as an agreement and compromise from various scientists regarding the number, name, and semantics of the interface functions.

A third technical issue discussed at the workshop was how to represent taxonomy data. Some researchers used a simple three level hierarchy for genus, family, and species. Other researchers had numerous tables and complex relationships to define the tree taxonomy. Some researchers were not interested in how their taxonomy, e.g., name changes and new discoveries, changed over time while other researchers wanted to capture every change made to the taxonomy.

There are currently taxonomic databases and solutions stored at botanical gardens and herbaria, including Tropicos, the Global Biodiversity Information Facility, and the International Plant Names Index (Tropicos) (GBIF) (IPNI). Their work is not tied to specific plot or survey data. Other work on taxonomies focus on representation issues, e.g., representing multiple, concurrent hierarchies (Raguenaud, Kennedy, and Barclay 1999). CTFS is currently building a taxonomic editor that is tied to plot data so that domain experts can make taxonomic changes and relate those changes to their plots.

AI Applications?

Most of the discussion at the CTFS workshop focused on conventional technology and approaches for sharing schemas, using data warehouses for integrating and sharing data, and representing taxonomies. We did not discuss innovative ways for scientists to integrate heterogeneous data sources and to share data, information, and knowledge.

For this reason, the AAAI workshop on Semantic Scientific Knowledge Integration intrigued us. We believe there is potential application of Semantic Web Services, Knowledge Grids, and Ontologies for sharing schemas, data, and data analyses in forest science. However, we have more questions than answers.

Potentially, Semantic Web Services could provide a range of automated tasks or services for searching, extracting, filtering, and integrating data (McIlraith, Son, and Zeng 2001) (Hendler 2001) (Cabral et al. 2004) (Berners-Lee, Hendler, and Lassila 2001). A botanist could request plot

data for "Rubiaceae.Alseis.blackiana", i.e., family, genus, and species respectively, in South America for a specific geographic area. After the request is made an automatic discovery service could go out and find and extract data from numerous web sites with data from Brazil, Colombia, etc. Each of these sites could invoke or execute web services to filter and extract subsets of data. Another service could collect and integrate all of the data. All of these services could be coordinated by a composition web service.

We are not sure whether any of these web services are feasible. Do any currently exist? Also we are not sure how web services handle data formatting. Do data standards already exist for transferring data? Are there new techniques for exchanging and combining data? Although a lot of work has been published on the Semantic Web, we'd like to better understand what functionality is real and can be applied to forest science.

A Knowledge Grid could provide advanced analytical capabilities in a cooperative environment, i.e., intelligent services and a problem solving environment to reason about data (Zhuge 2004). After large amounts of data are integrated, botanists will want to analyze the growth of certain species. A set of high level resources and services could be available to compare the presence and growth of a given species in different plots and understand factors like climate, topography, and soil on growth. Despite the hype about Knowledge Grid tools, we are not sure about their status and availability.

The work on ontologies also seem promising, however, we have a number of questions about their applicability. Can ontologies be used to help guide web services? Should ontologies represent information about the structure of the plot data? That is, can ontologies help support data exchanges between sites using different schemas? Can ontologies help handle incomplete or erroneous data? Can ontologies handle all of the different requirements associated with taxonomies?

Although two of the authors are computer scientists, the first author is the database designer for the STRI project, we would like to participate in this meeting as a representative of the scientific information community and learn more about current capabilities for Semantic Web Services, Knowledge Grids, and Ontologies. We would also like to generate discussion and better understand these technologies, i.e., whether they apply and whether they have advantages over conventional techniques.

References

Berners-Lee, T., Hendler, J., and Lassila, O. 2001. The Semantic Web, *Scientific American*, 284(5): 34-43.

Cabral, L., Domingue, J., Motta, E., Payne, T., and Hakimpour, F. 2004. Approaches to Semantic Web Services: An Overview and Comparison. In *The Semantic Web: Research and Applications*, C. Bussler, J. Davies, D. Fensel, and R. Studer eds. Berlin: Springer, 225-239.

CTFS website: www.ctfs.si.edu

GBIF website: www.data.gbif.org

Hendler, J. 2001. Agents and the Semantic Web, *IEEE Intelligent Systems*, 16(2): 30-37.

IPNI website: www.ipni.org

Kimball, R. 1996. *The Data Warehouse Toolkit*. New York: John Wiley & Sons.

McIlraith, S., Son, T.C., and Zeng, H. 2001. Semantic Web Services, *IEEE Intelligent Systems*, 16(2): 46-54.

Raguenaud, C., Kennedy, J., Barclay, P.J. 1999. Database support for taxonomy, Prometheus technical report #1, School of Computing, Napier University.

RAINFOR website: www.geog.leeds.ac.uk/projects/rainfor

SALVIAS website: www.salvias.net

Tropicos website: www.mobot.org/plantscience

US Forestry Inventory database website: www.fia.fs.fed.us

Vegbank website: www.vegbank.org

Zhuge, H. 2004. China's E-Science Knowledge Grid Environment, *IEEE Intelligent Systems*, 19(1): 13-17.