Web Service Support for Scientific Data Analysis

Martin J Kollingbaum*, Kun Cai**, Timothy J Norman, Derek Sleeman, and Wamberto Vasconcelos

Department of Computing Science, University of Aberdeen, AB24 3UE
{mkolling,kcai,tnorman,sleeman,wvasconc}@csd.abdn.ac.uk

1 Automated Scientific Analysis

The analysis of large data sets requires extensive computing resources and detailed knowledge about the analysis process itself. E-Science and Grid computing is an effort to provide computing resources for the scientific community and support collaborative research efforts. With the availability of integrated platforms that allow seamless communication and collaboration between researchers and access to distributed computing facilities, there is a strong case to add intelligent applications on top of these infrastructures that support and inform research activities and allow the automation of scientific workflows [1].

Providing an automated advice for scientists in their data analysis is not a new concern. Previous work has been conducted in the machine learning domain. Kodatoff et al. [3] and Sleeman et al. [4] describe an advisory system for this domain. The advisor uses a knowledge base that captures selection criteria for various machine learning algorithms. The central element of this system is a classification expert system, called Consultant, which interrogates the user about features of the learning task and recommends algorithms based on this information. Consultant shows the essential elements of an automated advisor in the machine-learning domain: information about the learning task (including characteristics of the data) is fed into a central knowledge base, which returns specific advice.

The Consultant expert system is a typical "single-vendor" monolithic tool. What we need today is an environment that accounts for the collaborative nature of research. The Web can be used to establish an advisory system that can be maintained by the research community. Consequently, the functionality of such a system is not established by a single group (organisation, company) but emerges from a collaborative effort of many people. This can be described as a "multi-vendor" scenario.

2 Web Services

Web service technology has been used to design a web-based scientific advisor. The move towards web services is driven by the requirement that it is main-

* Funded by the EPSRC IRC in Advanced Knowledge Technologies GR/NI 5764
** Partially funded by the EPSRC Master Training Program in E-Commerce Technology GR/N29709

tainable by the research community. In its essence, it shows the same elements as the original implementation: domain knowledge is captured as a set of rules comprising the knowledge base, which is used to advise on the selection of a specific algorithm. But it is not an advisor for any specific domain — the system represents a platform to which domain specifications can be submitted by users via the Web. Given such heuristics, a user in the role of a knowledge engineer can submit such a specification to the advisory system. The knowledge engineer engages in a dialogue with an administrator tool to specify algorithm selection heuristics in a convenient manner. This tool ultimately generates an OWL description of the engineer's specification, but in generating this description the tool may give guidance to the engineer by asking critical questions such as: "is decision criteria x derivable from the data set or must it be user-specified?". Once this design is complete, the administrator tool automatically generates a set of Jess[1] rules that provide a procedural encoding of the domain heuristics. This domain knowledge is subsequently available for any user seeking advice. Figure 1 shows the essential elements of this system with the two basic roles for users. The so-called "administrator" is responsible for the deployment of new domain-specific knowledge bases. These are wrapped by so-called "broker" services, which provide the advisory service.

Future research has to take into account the open nature of such a web-based advisory infrastructure. Because of its openness, its functionality emerges from the concerted effort of the research community. It also offers possibilities for commercial vendors to contribute descriptions and knowledge about products implementing scientific analysis or to access various sources that can provide data. With scientists using third-party data for their analysis or with commercial interests involved, problems of privacy, charging models, trust in and reputation of services have to be tackled [2]. Using agent technology in this advisory scenario can provide the necessary capability to negotiate contracts for the usage of specific services or to perform analysis tasks without disclosing privacy issues of data sets.

References

1. D. De Roure, N.R. Jennings, and N.R. Shadbolt. The Semantic Grid: Past, Present and Future. In *to appear*, 2005.
2. I. Foster, N.R. Jennings, and C. Kesselman. Brain Meets Brawn: Why Grid and Agents Need Each Other. *Proc. 3rd international Conference on Autonomous Agents and Multi-Agent Systems AAMAS 2004*, pages 8–15, 2004.
3. Y. Kodratoff, D. Sleeman, M. Uszynski, K. Causse, and S. Craw. Building a Machine Learning Toolbox. In L. Steels and B. Lepape, editors, *Enhancing the Knowledge Engineering Process*, pages 81–108. Elsevier Science Publishers, 1992.
4. D. Sleeman, M. Rissakis, S. Craw, N. Graner, and S. Sharma. Consultant-2: pre- and post-processing of Machine Learning Applications. *International Journal of Human-Computer Studies*, 43:43–63, 1995.

[1] http://herzberg.ca.sandia.gov/jess