KNOWLEDGE ACQUISITION FROM AUTONOMOUS, DISTRIBUTED, SEMANTICALLY HETEROGENEOUS DATA SOURCES

Doina Caragea[1,4], Adrian Silvescu[1,4], Jyotishman Pathak[1,4], Jie Bao[1,4], Carson Andorf[1,3,4], Changhui Yan[1,3,4], Drena Dobbs[2,3,4] and Vasant Honavar[1,2,3,4]

[1]Artificial Intelligence Research Laboratory, Department of Computer Science, 226 Atanasoff Hall
[2]Department of Genetics, Development and Cell Biology, 1210, Molecular Biology Building
[3]Bioinformatics and Computational Biology Program, 2014, Molecular Biology Building
[4]Computational Intelligence, Learning, and Discovery Program, 214 Atanasoff Hall
Iowa State University Ames, IA 50011

honavar@cs.iastate.edu

Ongoing transformation of biology from a data-poor science into an increasingly data-rich science, with the attendant increase in the number, size, and diversity of sources of data (e.g., protein sequences, structures, expression patterns, interactions) offer unprecedented, and as yet, largely unrealized opportunities for large-scale collaborative discovery in a number of areas including characterization of macromolecular sequence-structure-function relationships, discovery of complex genetic regulatory networks, etc.

Given the large number, autonomous nature and the size of the relevant data sources, gathering all of the data in a centralized location is generally neither desirable nor feasible. Hence, there is a need for methods to perform the necessary analysis of data where the data and the computational resources are available and transmit the results of analysis (knowledge acquired from the data) to where they are needed. More importantly, data sources developed by autonomous individuals or groups differ with respect to their ontological commitments (that is, assumptions concerning the *objects* that exist in the *world*, the *properties* or *attributes* of the objects, the possible *values* of attributes, and their *intended meaning*). Therefore, semantic differences among autonomous data sources are simply unavoidable. Because data sources that are created for use in one context often find use in other contexts or applications and because users often need to analyze data in different contexts from different perspectives, there is no single privileged ontology that can serve all users, or for that matter, even a single user, in every context. Effective use of multiple sources of data in a given context requires flexible approaches to reconciling such semantic differences from the user's point of view.

To address the information integration and knowledge acquisition needs of collaborative scientific discovery, we have designed INDUS (Intelligent Data Understanding System), a *federated*, *query-centric* system for knowledge acquisition from distributed, semantically heterogeneous data (See **Figure 1**). INDUS employs ontologies and inter-ontology mappings, to enable a user or an application to view a collection of physically distributed, autonomous, semantically heterogeneous data sources (regardless of location, internal structure and query interfaces) as though they were a collection of tables structured according to an ontology supplied by the user[1]. This allows INDUS to answer user queries against distributed, semantically heterogeneous data sources without the need for a centralized data warehouse or a common global ontology.

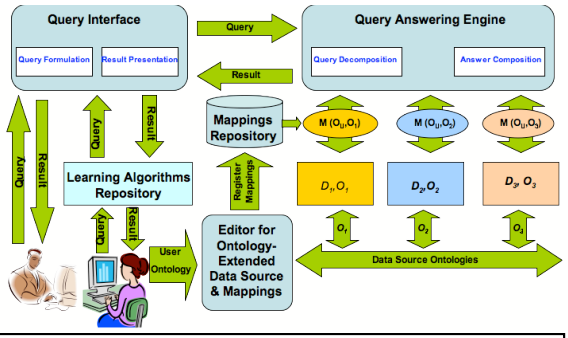

Figure 1: INDUS: a system that supports knowledge acquisition from semantically heterogeneous distributed

We used INDUS framework to design algorithms for learning probabilistic models for predicting GO functional classification of a protein based on training sequences that are distributed among semantically heterogeneous data sources (SWISSPROT and MIPS). Mappings such as EC2GO and MIPS2GO were used to resolve the semantic differences between these data sources when answering queries posed by the learning algorithms. Our results show that INDUS can be successfully used for integrative analysis of data from multiple sources needed for collaborative discovery in computational biology.

Acknowledgements: This work was funded in part by grants from the National Science Foundation (IIS 0219699) and the National Institutes of Health (GM 066387).

---

[1] Caragea, D., Pathak, J., and Honavar, V. (2004). Learning Classifiers from Semantically Heterogeneous Data. In: Proceedings of the Third International Conference on Ontologies, DataBases and Applications of Semantics for Large Scale Information Systems (ODBASE'04), Lecture Notes in Computer Science Vol. 3291 pp. 963-980, Berlin: Springer-Verlag.