INDUS: A System for Information Integration and Knowledge Acquisition from Autonomous, Distributed, and Semantically Heterogeneous Data Sources

Jyotishman Pathak[1,4], Jie Bao[1,4], Doina Caragea[1,4], Adrian Silvescu[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

INDUS (Intelligent Data Understanding System) is a *federated*, *query-centric* system for knowledge acquisition from distributed semantically heterogeneous data sources that employs ontologies (controlled vocabularies of domain specific terms, and relationships among terms) and inter-ontology mappings, to enable a user to view a collection of such 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. INDUS and the associated collection of software tools

(a) Support editing of ontologies and specification of semantic relationships between ontologies (using inter-ontology mappings [Bao and Honavar, 2004]) by users with some familiarity with the data sources, using a graphical user interface.

(b) Enable users to query distributed, semantically heterogeneous data and retrieve and manipulate results in a fashion that respects the user-imposed semantic relationships between different sources of data [Caragea et al., 2004b].

(c) Construction of predictive classifiers from semantically heterogeneous distributed data sources without having to assemble all of the data at a central location [Caragea et al., 2004a; Caragea et al., 2004b]. This is achieved by decomposing the task of learning from data into an information extraction task, that formulates and sends a statistical query to a data source, and a hypothesis generation task, that uses the resulting statistic to modify a partially constructed hypothesis (and further invokes the information extraction component as needed).

INDUS framework has been successfully used 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) [Andorf et al., 2004]. Mappings such as EC2GO and MIPS2GO are used to resolve the semantic differences between these data sources when answering queries posed by the learning algorithms.

Major applications: Data Integration and Knowledge Acquisition from Semantically Heterogeneous Biological Data.
Platform: Runs on any platform supporting JDK 1.4 or above. User should have a RDBMS (e.g., Oracle, Postgre SQL) installed and the required JDBC libraries.
Availability: Open source software issued under the GNU General Public License.
Where: http://www.cild.iastate.edu/software.htm (will be made available by the end of March, 2005).

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

References

1. Andorf, C., Silvescu, A., Dobbs, D. and Honavar, V. (2004). Learning Classifiers for Assigning Protein Sequences to Gene Ontology Functional Families. In: *Fifth International Conference on Knowledge Based Computer Systems (KBCS 2004)*. India.

2. Bao, J. and Honavar, V. (2004). Collaborative ontology building with wiki@nt - a multi-agent based ontology building environment. In: Proc. of 3rd International Workshop on Evaluation of Ontology based Tools, located at the 3rd International Semantic Web Conference ISWC 2004, 8th November 2004, Hiroshima, Japan.

3. Caragea, D., Silvescu, A., and Honavar, V. (2004a). A Framework for Learning from Distributed Data Using Sufficient Statistics and its Application to Learning Decision Trees. *International Journal of Hybrid Intelligent Systems*. Vol. 1, No. 2. Invited Paper.

4. Caragea, D., Pathak, J., and Honavar, V. (2004b). 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), October 25-29, 2004, Agia Napa, Cyprus.

5. Reinoso-Castillo, J., Silvescu, A., Caragea, D., Pathak, J., and Honavar, V. (2003) A Federated Query Centric Approach to Information Extraction and Integration from Heterogeneous, Distributed, and Autonomous Data Sources. In: The 2003 IEEE International Conference on Information Reuse and Integration, October 27-29, 2003, Las Vegas, USA. IEEE Press. pp. 183-191.