Toward Semantic Interoperability of Heterogeneous Biological Data Sources

Sudha Ram

Eller Professor of MIS, Director, Advanced Database Research Group,
Eller College of Management, University of Arizona,
Tucson, AZ 85721
ram@eller.arizona.edu
http://vishnu.eller.arizona.edu/ram

Genomic researchers use a number of heterogeneous data sources including nucleotides, protein sequences, 3-D Protein structures, taxonomies, and research publications such as MEDLINE. This research aims to discover as much biological knowledge as possible about the properties and functions of the structures such as DNA sequences and protein structures and to explore the connections among all the data, so that the knowledge can be used to improve human lives. Currently it is very difficult to connect all of these data sources seamlessly unless all the data is transformed into a common format with an id connecting all of them. The state-of-the-art facilities for searching these data sources provide interfaces through which scientists can access multiple databases. Most of these searches are primarily text-based, requiring users to specify keywords using which the systems search through each individual data source and returns results. The user is then required to create the connections between the results from each source. This is a major problem because researchers do not always know how to create these connections. To solve this problem we propose a semantics-based mechanism for automatically linking and connecting the various data sources. Our approach is based on a model that explicitly captures the semantics of the heterogeneous data sources and makes them available for searching. In this talk I will discuss issues related to capturing the semantics of biological data and using these semantics to automate the integration of diverse heterogeneous sources.

O. Pastor and J. Falcão e Cunha (Eds.): CAiSE 2005, LNCS 3520, p. 32, 2005.