Example Project: Genomic Knowledge Inference

Project Summary

- This research project investigates the information technology and infrastructure to better support automated and on-demand knowledge inference from geographically distributed biomedical databases. Specifically, the project develops efficient techniques to extract and analyze knowledge from a distributed Unified Medical Language System (UMLS). The underlying information infrastructure is a campus-wide computational and data Grid. Related research work is part of the Bio-Grid research initiatives currently conducted at the University of Connecticut.

- This research work involves a few tasks, including (1) the development of a distributed semantic network system, based on a task-based and message-driven model to exploit both task and data parallelism while processing queries; (2) the parallelization of the inference engine to speed-up the query processing; and (3) automated data migration among the distributed knowledge bases to maximize the storage utilization rate. The current information infrastructure, as a test-bed, of this project is a campus-wide computational and data Grid. Participating sites of this infrastructure include the Schools of Engineering, Public Health and Medicine at the University of Connecticut. Note that the Grid represents a rapidly emerging and expanding technology that allows geographically distributed resources (CPU cycles, data storage, sensors, visualization devices, and a wide variety of Internet-ready instruments), which are under distinct control, to be linked together in a transparent fashion. The aggregate computing power, data storage, network bandwidth, as well as the user friendliness have rendered the Grid a prosperous infrastructure in support of automated processing of distributed information. Our system features multi-threading and task migration to support communication latency hiding and load balancing, respectively.

- Grid-enabling the UMLS is part of the Bio-Grid initiatives currently conducted at the University of Connecticut. The initiatives intend to establish a campus-wide information infrastructure in support of automated and on-demand distributed knowledge exploration in biomedical informatics, which is increasingly needed by life-science research work. Application areas in biomedicine include the epidemiological studies and medical imaging, which produce tremendous amount of data that are usually geographically distributed among hospitals, clinics, research labs, and radiology centers, etc. For research, training or clinical purposes, physicians and researchers often need to consult and analyze data from distributed sites. Thus, an infrastructure supporting on-demand and automated information extraction and reasoning will provide significant convenience.

Components for Summer Students

- Students participating in this project learn about semantic networks, the UMLS, biomedical knowledge representation and basic concepts of distributed knowledge reasoning. Students also involve in the design of the distributed UMLS. The design experience of the distributed UMLS, particularly the task model for cooperative inference and the layered architecture for the host and slave systems, complies with the health data management (HDM) unit, a core component under development, in which the metadata management capabilities can easily be designed. The HDM design for secure retrieval of sensitive bio/medical/health data was motivated by our research on a distributed UMLS. Students also learn the overall system design concept of the distributed UMLS.

Statistics (2010)

- Students demographics: 10 per year
  - Total Under-Represented: 70 %
  - Under-Represented minority: 20 %
  - Female: 60 %
  - From undergraduate school: 50 %
  - From diverse institution: 90 %
  - Total number of mentors: 6
  - Years of operation: 2008 to 2010
- Example Research topics: Modern Computing Infrastructure, Protein Function Studies, Genomic Knowledge Inference, Bioethics in Health-Grids, caBIG and BIRN as Case Studies.

Student Publications

- To be reported.