Scalable Multi-Source Astronomy Data Mining in Distributed, Peer-to-Peer Environments

Kamalika Das, Wesley Griffin, Hillol Kargupta, University of Maryland, Baltimore County
Chris Giannella, Loyola College
Kirk Borne, George Mason University

Design, implementation, and archival of very large sky surveys are playing an increasingly important role in today's astronomy research. Current projects such as GALEX All-Sky Survey and future ones such as WISE All-Sky Survey are destined to produce enormous catalogs of astronomical sources. The Large Synoptic Survey Telescope is supposed stream in large volumes of data at a high rate. It is this virtual collection of gigabyte, terabyte, and (eventually) petabyte catalogs and streams that will enable remarkable new scientific discoveries through the integration and cross-correlation of data across these multiple survey dimensions. However, this will be difficult to achieve without a computational backbone that includes support for queries and data mining across distributed virtual tables of de-centralized, joined, and integrated sky survey catalogs. Moreover, use of local data management systems such as MyDB, MySpace in AstroGrid, and Grid Bricks for storing and managing user's local data is becoming increasingly popular. This is opening up the possibility of constructing Peer-to-Peer (P2P) networks for data sharing and mining.

This research is exploring the possibility of using distributed and P2P data mining technology for exploratory astronomy from data integrated and cross-correlated across these multiple sky surveys. It is considering several scientific problems in order to illustrate the possibilities. For example, we are exploring classical fundamental plane problem in a new light which is trying to answer some of the following questions: How does local galactic density relate to galactic fundamental plane structure? Does the fundamental plane structure of galaxies in low density regions differ from that of galaxies in high density regions? Since the attributes which define the fundamental plane span two data repositories SDSS and 2MASS instead of one, we focus on cross-matching them available individually through the NVO. We are using distributed data mining algorithms to analyze this data distributed over a large number of nodes.

Distributed data mining techniques will not require scientists to download massive chunks of data for scientific discovery and will enable them to use distributed database queries across distributed virtual tables of de-centralized, joined and integrated sky survey catalogs. This will make the existing client-server-based astronomy data services richer by providing the power of distributed and P2P data mining technology. This research will also contribute toward scaling up mining of astronomy data in a cloud/grid computing environment that are gradually becoming popular within the scientific community for efficient processing of very large datasets. For the galactic density-variation problem, the major challenges in distributed data mining are (i) how to create a global ordering of the galaxies based on density without looking at the entire data, and (ii) how to perform a distributed eigen analysis of the data using communication-efficient local algorithms. Our current contributions in this area include development of approximate algorithms offering considerably lower communication cost than centralization. We envision this study as only the beginning of a longer series of studies designed to address astrophysical questions that could not be as readily addressed using individual standalone databases or computing resources.