Astronomical Data Analysis Software and Systems IX
ASP Conference Series, Vol. 216, 2000
N. Manset, C. Veillet, and D. Crabtree, eds.

Scientific Data Mining

P. Dowler, D. Schade, R. Zingle, D. Durand, S. Gaudet

Canadian Astronomy Data Centre, National Research Council of Canada, Victoria, BC

Abstract. We have started an innovative project to develop scientific data mining technology and use it to deliver powerful research and analysis tools to the astronomical community. Our goal is to provide access to multiple science archives with support for exploratory searches and general ad-hoc querying capabilities.

1. Introduction

The CADC has undertaken an innovative and ambitious project to develop a data-mining system which will enable astronomers to search and analyze the vast amount of available scientific data in a structured and efficient manner (Schade et al. 2000). The data-mining project involves two new and exciting avenues of research. The first is the development of a science archive which stores both pixel data and scientific results in a highly cross-referenced database which supports ad-hoc querying by users. This archive is primarily an extragalactic object archive based on the results of large collaborations, surveys, and other major astronomical projects. Second, we are developing a multi-tier (client-server) system to support efficient exploration, querying, and analysis of the science archive content, distributed processing of pixel data to create new scientific results, and (eventually) uniform access to other information services (external sites which provide interesting astronomical data, catalogs, preprints, electronic publications, etc.).

The central concept in the CADC data-mining architecture is that the user specifies the information required to answer their scientific question and the data-mining system acquires all the available information. The acquisition of information may involve queries to the science archive, access to external information services, and server-side or client-side processing of raw data or intermediate results. We are developing client software to aid the users interaction with the data mining system which supports the building of information requests, receives and organizes the requested information, and provides a variety of visualization and analysis tools to aid in interpreting the results. We also intend to support third party tools through a published interface that uses XML to format both information requests and the results. This XML interface would be made up of two parts: a DTD which specifies a generic scientific querying language and a DTD which specifies a generic scientific data structure.

2. Data Archive vs. Science Archive

Most of the online resources available to the astronomical community are data archives. Users of these services typically search according to observational parameters (telescope, type of observation, coordinates, intended targets, exposure time, etc.) and retrieve raw or calibrated image data which is then processed with a variety of standard analysis tools. In addition to traditional data archives, many astronomical catalogs can be searched online, but it is generally difficult or impossible to cross reference different catalogs or search several catalogs simultaneously. Some new projects have made encouraging progress in improving catalog access and integrating catalog and data access (Vizier and Aladin come to mind).

A science archive stores the results of data analysis - the scientific measurements - rather than the data. For the astronomical community, an ideal science archive would store all the known properties of all known objects in the universe. Users of such a service would be able to search according to scientific parameters (magnitude, redshift, spectral indices, morphological type of galaxies, proper motion of stars, etc.) and find the set of objects and parameters that will help answer the scientific question at hand. To be useful as a tool for actually doing science, a science archive needs maintain links between the scientific results and the image data and processing tasks which created them. This support for "drill-down" is critical because it allows scientific users to see exactly where a result comes from and verify its reliability and importance.

3. Exploration and Querying

The science archive is only useful if users are able to explore it and discover what information is available in an intuitive and dynamic fashion. At the simplest level, this means the user is able to discover easily the types of objects in the archive and the list of properties that might be available for those objects. Our model for exploration and querying is that users define the set of objects which they are interested by constraining the values of their scientific properties and then submit the set definition (i.e. the constraints) along with the desired level of detail (LOD) as a request. The constraints define the set and the LOD defines the view of the set that is returned as a result. The lowest LOD returns the number of objects in the set. The next LOD returns the range of values for each property of interest in the set. The third LOD returns the distribution of values for each property (histograms), and the highest LOD returns the actual values. The user can iteratively refine the constraints and alter the LOD until they find the set of objects and properties that suits their needs.

We have developed a robust but simple object-oriented (OO) constraint system to encapsulate conditions on set membership. This design makes it straightforward to implement a graphical user interface (GUI) to create and edit constraints. In addition, since the constraint system and LOD make up a hierarchical data structure, it is easy to serialize them using XML and maintain the structure.

4. Managing Content

The science archive stores millions (billions) of objects with hundreds of possible properties per object, but this content is likely to be extremely inhomogeneous with many or even most object-property pairs missing. We assume this will always be the case because it is not plausible to keep up with the task of applying every analysis tool to every bit of observational data. Furthermore, it would hide any meaningful results in a vast wasteland of noise. Instead, the servers-side must deliver a manageable, (nearly) homogeneous subset of objects and properties. The client-side then enables the user to visualize, verify, and analyze the results.

We have developed a generic object-oriented data structure to store and manipulate scientific content (sets, histograms, tables, images, etc.) and provide client-side software with the ability to explore, visualize, overlay, and merge query results in a sophisticated fashion. This scientific data structure includes both the content and any meta-data which describes the content together to allow users to fully evaluate any scientific interpretations that they make. This meta-data includes things like the explicit selection effects (the constraints that define the set), implicit selection effects (derived from the selection effects on source data), and statistical measures of the completeness and reliability of the content.

5. Data Mining

Our research into existing data mining technologies has shown that mining scientific data involves complexity not found in the traditional business world. Traditional data mining is designed to find unexpected relationships or patterns in the source data. The dimensionality of the source data is typically modest and the content is quite homogeneous (you do not have to deal with sparse data sets). In addition, the data itself is usually discrete (integer data) and there is not a significant amount of processing involved in getting it from underlying raw data. In the scientific community, however, we deal with inhomogeneous (sparse) datasets with high dimensionality. This scientific data is derived from extremely large sets of raw data through after extensive processing. On top of all that, the results we are trying to mine are floating point values, sometimes with known error limits, which only approximate the true value of the parameter.

Rather than attempt to implement an artificial scientist, we have opted to design and implement an infrastructure that provides interactive and iterative exploration and analysis. Our system will allow researchers to actively do science online rather than just gather information for later analysis.

6. System Architecture

Our prototype system is a 4-tier system developed in the Java language. The client-side is currently implemented as a Java application which allows the user to construct queries (constraint set + LOD) and submit them to the server and which then receives results as they become available. On the server-side we have a tier which supports client connections, authentication, and state (session management). The session server constructs data requests and forwards them to the data server for processing. The data server sits between the back-end services and the session server and performs two tasks. First, it determines which services can supply content to fill the data request and submits the request to them. Second, it receives content from the various back-end services, performs necessary merging, cross-correlation, or post processing, and then sends the completed results back to the session server. The back-end of the system is made up of 3 components which provide different data services: a database, a processing subsystem, and an external archive/catalog interface. The database holds the existing science archive content and responds by executing queries built from the constraint set. The processing subsystem has access to data analysis software and image data and applies these tools to images to produce new scientific results. These results are sent to the data server and, in some case, also ingested into the database for future use. The external archive and catalog interface requests content from other astronomy sites and online resources.

7. Summary

Our goal is to provide a sophisticated scientific data mining infrastructure which enables astronomers to take utmost advantage of the immense scale of current and future data and science archives. We are implementing a multi-tier system which provides uniform access to back-end services such as querying our database archives, online processing of raw data, and the querying of other resources in the community at large.

Acknowledgments. This project has been funded primarily through a special grant from Dr. J. Lyrette (Vice President, National Research Council). The entire CADC staff provide input into various aspects, both technical and otherwise, for this project.

References

Schade, D. et al. 2000, this volume, 215