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

A Data Mining Model for Astronomy

D. Schade, P. Dowler, R. Zingle, D. Durand, S. Gaudet, N. Hill, S. Jaeger, D. Bohlender

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

Abstract. The content of astronomical data archives has grown dramatically during the past decade whereas the ability of scientists to access that content remains primitive and ineffective. Astronomical datasets are large, distributed, and contain a wide variety of data types—from raw pixels to fully-derived tables of scientific measurements. These datasets are ideal for the development and deployment of Data Mining tools and systems. As a first step toward defining a Data Mining system, we describe some components of the infrastructure that is required to enable effective Data Mining in astronomy and we outline some of the major obstacles to the production of a powerful Data Mining system.

1. Introduction

The list of archival data resources available to researchers in astronomy is impressive. For example, data is available at all wavelengths from the X-ray (heasarc.gsfc.nasa.gov) through the optical (archive.stsci.edu, cadcwww.dao.nrc.ca) to the infrared (www.ipac.caltech.edu), sub-millimetre (cadcwww.dao.nrc.ca/jcmt) and radio (sundog.stsci.edu) regimes. A compendium of resources is shown at the Astrobrowse home page (guinan.gsfc.nasa.gov/ab). These facilities combine the data with the relevant scientific expertise needed to support their effective archival use.

The most impressive scientific gains from astronomy data archives will be achieved by combining data from many sources and performing joint analyses of the resulting datasets. However, at the present time it is difficult and time consuming to locate the necessary components of such a combined dataset, to ascertain if these components are suitable, to retrieve the relevant datasets, and to process them into a form suitable for joint analysis. Excellent science is lost because of these difficulties.

What type of information management systems need to be in place in order to create an environment where many more archival research projects are successfully undertaken? We have outlined a few of the major components of such a Data Mining system and sketched their requirements and associated difficulties.

2. Scientific Language Query Tools OR Tools to Explore and Characterize the Database

Ideally, queries should be expressed in scientific language rather than data-oriented terminology. Alternatively, one could interactively explore a particular database or collection of databases as the initial phase of a Data Mining project.

An example query might be: "Give me the available images of Abell clusters at $z > 0.02$ and the redshifts available in each cluster. Only show those clusters with available X-ray data." This is a very modest request and such data exists on the web, yet this query is impossible to execute on present systems. It would take several hours to do a moderately thorough search of the appropriate database sites and a many more hours to retrieve and construct a usable sample. The time required to execute effective searches for scientifically interesting datasets quickly reaches the point where it inhibits the exploratory work which is necessary at this point in archive development.

An easier but more limited approach to this problem is to provide tools to explore the contents of specific databases. One could examine plots and histograms of the database content in order to understand what is feasible and what is not. This cannot at present be done on major archives and implementation would be limited to local databases until cross-archive access is improved.

3. Query Constructor

This component receives the query (which may have been made in scientific language) and knows where to direct its requests for relevant data. It needs to know how to format its queries and how to understand the results that are returned. A major part of the difficulty here is related to the non-standard form for queries and results from remote archives.

For example, a query construction tool needs to translate science terms (bright QSO) into references to resources (the Palomar-Green Survey, Burbidge catalog, etc.). The returned results need to be complete and reliable in some well-defined way. Although many archives allow querying on object type, for example "GALAXY CLUSTER", there is no guarantee that the results are complete. The association of keywords with data in the HST archive is a step in the right direction but completeness and reliability of such a system needs to be verified. Otherwise, the value of such a system is small.

4. Access to Local and Distributed Data Archives

Current access to archives is primitive. Queries key on position in the sky or object name. *Archive contents need to be indexed according to scientific criteria.* Cross archive queries need to be supported much more effectively than they are at present.

The ASTROBROWSE tool (guinan.gsfc.nasa.gov/ab/) allows a user to search the major archive sites for data on a particular sky position (or a particular object via a name resolver). Queries are very basic and the returned results can be numerous, overwhelming, and difficult to evaluate. Extending this framework should a priority among the archive community.

5. Existence of and Access to Metadata

A key missing element in current archives is the packaging of relevant metadata together with the data themselves. Metadata is information that describes or defines the data themselves. As an important example, the selection function for each piece of data needs to be available and to accompany the data. This and other forms of metadata are needed to feed into the following phases of the Data Mining process.

The characterization of the selection function is a difficult but important problem. A scientist attempting to construct a survey sample which is "unbiased" with respect to some attribute must have access to a reliable description of the selection criteria which led to the choice of a particular field for observation. Without this information, hidden biases will creep into the constructed sample.

Current archives do not generally have sufficient information about the selection functions operating on their content. Large surveys have simple and well-defined selection functions and thus are much more useful for the purposes of Data Mining than are general archives of data from many individual proposals.

Other important missing metadata for ground-based observatories include weather monitoring, observation logs, and logging of other significant observatory events.

6. Processing

A Data Mining system would require processing at various stages. Ingestion might require fundamental calibration and would require processing to integrate new data into the existing system (generating new indices, object cross-identification). Later, processing may be necessary in order to fulfill a query where both the input data and the algorithm exist but where processing has not yet been carried out. These new results may be ingested into the database. Processing may also be needed where heterogeneous datasets have been retrieved and these need to be homogenized if order for analysis to proceed.

Processing is part of the selection and preparation of data but it must be recognized that not all possible data processing can be done ahead of time and the results stored. For example, it is not possible to anticipate all useful combinations of data and, furthermore, new data ingestion creates new opportunities for joint analysis.

7. Analysis Tools

This is the true Data Mining phase. Various algorithms can be applied including machine learning or automated discovery tools. However, all of the conventional scientific visualization and analysis tools are also applicable at this stage.

The result of the analysis phase may be the establishment of the final results or, more frequently, refinement of the query or the scientific question itself and reiteration.

Much of what is described here is really infrastructure that is needed to produce a Data Warehouse that will then enable Data Mining.

8. Generating Content for Data Mining Systems

Our consideration of the problem of Data Mining has led to several conclusions. The most profound one is the following. If one sets as a goal the creation of a highly-effective information-management environment for astronomy (whether it is called Data Mining or something else), then impacts are created on all phases of the process of doing astronomy. These impacts are not trivial. For example, information that can be carried in the head of a single proposer-observer-analyst-interpreter-author as part of the current model of astronomical research, needs to be collected in machine-readable form at all stages (proposal, observation reduction) of the process. Thus new obligations and responsibilities for information collection are created for observatories.

A second clear conclusion is that surveys represent by far the best content for Data Mining. The advantages of surveys include a clean and comprehensible selection function, homogeneity, and size. New surveys such as the Sloan Digital Sky Survey will join existing survey material in many different energy bands to produce a vast increase in the power of archival analyses.

A third conclusion is that much of the archive content of ground-based observatories that exists in 1999 will be usable only in a limited way in sophisticated Data Mining systems because a great deal of information has been irretrievably lost.

The designers of new observatories need to consider good information management one of the key components of their systems which, after all, produce nothing except information. The next generation of ground and space-based observatories will show improvement but it may not be until the following generation that the output of ground-based observatories can be used as input for sophisticated Data Mining systems.

9. Summary

- Data need to be indexed according to scientific criteria in order to enable better querying;
- Effective means of cross-archive access are needed;
- Surveys with clearly-defined selection criteria are the most suitable input for Data Mining;
- The analysis of surveys is a data-driven process and is thus amenable to Data Mining procedures;
- Many existing astronomical data archives are unsuitable for Data Mining because key pieces of metadata are missing. Retro-fitting is only possible in some cases;
- The desire for the capability to do Data Mining leads to new requirements at all phases of the data collection process. Proposals, planning and execution of observations, logging procedures, data headers, pipeline processing, and archiving and distribution procedures all need to be re-thought and re-designed to be compatible with Data Mining requirements.