Data-Driven Discovery through e-Science Technologies

Kirk D. Borne
George Mason University, QSS Group Inc., and NASA-GSFC
Kirk.borne@gsfc.nasa.gov

Abstract

Future space missions and science programs will be massive data producers. The technology to produce large data volumes must be matched by technologies to process, analyze, and make use of the data flood, in order to reap the maximum engineering benefit and scientific return from those technology investments. In particular, the integration of data from multiple sources will be standard practice, both for operational decision-making and for scientific decision-making. We describe the application of the emerging e-Science paradigm and its related technologies to data-driven discovery in space missions of the future.

1. Introduction

In the data-driven world of science, where the impending data avalanche from future missions is likely to overwhelm our analysis and decision-support systems, it is imperative that we find solutions that extract as much knowledge from the data streams in real-time as possible. Otherwise, critical discoveries (scientific and spacecraft safety-related) may escape timely detection or else go completely undiscovered. In particular, in robotic exploration systems and other autonomous science mission planning systems, if a critical event or discovery is identified, then how does this information percolate up to the workflow, processes, and decision-makers? This is especially time-critical if the event requires some immediate follow-up scientific measurement or engineering maneuver.

The application of an e-Science paradigm, including distributed computing, Web Services, Semantic Web technologies, and machine learning algorithms will enhance the scientific return and knowledge-building capabilities of future space missions [1]. This will enable scientists and engineers to address data-intensive problems that would not otherwise be manageable. This will permit future space missions to make use of larger data volumes in the discovery and decision-making processes than is currently possible. We describe e-Science solutions to the challenges of future space missions within the context of the exponential growth in mission data streams.

2. e-Science

E-Science refers to the internet-enabled sharing of distributed data, information, computational resources, and team knowledge for the advancement of science. The technologies that enable this distributed sharing of resources are the same technologies that enable the equivalent paradigms of e-Commerce, e-Business, and e-Government. Machine-to-machine protocols and standards are used for passing messages between processes (e.g., queries and responses), for describing services and resources (e.g., data, software packages, computational environments), for discovery of these resources (e.g., through open registries), and for integrating data and responses from distributed systems. These technologies enable web portals to perform their function: one-stop shopping for airline tickets, for hotels, for books, or for data. E-Science thus builds upon and takes advantage of technologies that are built to satisfy economic and functional requirements. The scientific user requirements of e-Science are sufficiently parallel to those of e-gov and e-biz that technology transfer is relatively straightforward. Typical user requirements on modern internet-based science data systems include: find the right data right now; one-stop shopping for all data needs; integrate data from multiple sources with minimal human intervention; transparent user access to heterogeneous databases and query systems; metadata-assisted data fusion; identify the most important and/or unique events within a massive data collection; and knowledge discovery in databases (KDD).

Features of e-Science applications that address space mission challenges include: (1) event/anomaly detection and classification in real-time data streams (through machine learning); (2) data assimilation in real-time models; (3) information/knowledge representation and sharing (through ontologies); and

2nd IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT'06)
0-7695-2644-6/06 $20.00 © 2006 IEEE

(4) dynamic data-driven steering of mission measurement systems. We will address aspects of these, but first we illustrate an e-Science implementation: in the field of astronomical research.

2.1 e-Science implementation for astronomy

The development of models to describe and understand scientific phenomena has historically proceeded at a pace driven by the acquisition of new data. The more we know, then the more we are driven to tweak, or to augment, or to revolutionize our models and our understanding. This data-driven modeling and discovery linkage is entering a new paradigm.

The acquisition of scientific data in all disciplines is now accelerating and causing a nearly insurmountable data avalanche. The problem likely will get more extreme before it gets better. In astronomy in particular, rapid advances in three technology areas (telescopes, detectors, and computation) have continued unabated, with no slowdown in sight.

This accelerated advance in data generation capabilities will require novel, increasingly automated, and increasingly more effective knowledge discovery and modeling systems. Humans will be in the loop, but they will not comprise a major portion of the loop. In fact, enabling a telescope (or space mission) to act autonomously will lead to more efficient and effective science operations.

One of the challenges posed by the data avalanche is knowing what data exist that can be applied to a particular research problem or operational scenario. This challenge area can be represented by the following series of questions:

(Q1) How does one know whether the data exist?

(Q2) How does one find the data?

(Q3) How does one integrate those data with other data, models, and simulations?

(Q4) How does one extract new knowledge and understanding from the flood of new data and models?

(Q5) If a new discovery is made during this process, then how does one generate a request back to the data producer (e.g., deep space mission) to acquire some additional or new type of supporting data?

In the era of "small data," these questions were not a significant problem. In the new era of "insurmountable data," these questions represent valid and challenging research topics. As an illustration of this challenge area, we describe how the astronomy community is tackling these questions. In particular, the astronomy community has embarked on a grand information technology program, to describe and unify all astronomical data resources worldwide. This global interoperable virtual data system is referred to as the National Virtual Observatory (NVO at {www.us-vo.org}) in the U.S., or more simply the Virtual Observatory (VO). Within the international research community, the effort is steered by the International Virtual Observatory Alliance ({www.ivoa.net}).

This grand vision encompasses more than a collection of data collections. The result of these initiatives will represent a significant evolution in the way that astrophysical research (both observational and theoretical research) is conducted in the new millennium [2]. The VO effort will enable discovery (Q1), access (Q2), and integration (Q3) of data, tools, and information resources across all astronomy observatories, archives, data centers, and individual projects worldwide [3].

However, it remains outside the scope of the VO projects to generate knowledge, new models, and scientific understanding from the huge data volumes flowing from the largest sky survey projects (Q4) [1] [2]. Further removed from the VO still is the ensuing feedback and impact of the potentially exponential growth in new scientific models on those telescope instrument operations (Q5). The nature of the scientific method is to ask questions, collect data, get answers, ask new questions, collect new data, and so on.

The problem therefore is this: astronomy researchers will soon (if not already) lose the ability to keep up with any of these things: the data flood, the scientific discoveries buried within, the development of new models of those phenomena, and the resulting new data-driven modifications to the observing strategies of the telescope-detector measurement systems (in order to collect new data to answer the new questions raised by the new models).

What is the solution? One solution is to empower the data-collection systems to act autonomously to address those 5 questions: search for relevant data, access the data, integrate the data into the system, learn from the combined data stream, and then respond (e.g., schedule new observations, or new mission operations).

3. Machine learning

Machine learning refers to the application of algorithms that learn from experience. In the context of NASA space missions, this means that the system learns from the data that are being collected. The data may be scientific data, engineering/instrumental data, or telemetry/spacecraft data. The system that is

learning may be the scientific data acquisition system, the operations/scheduling system, or the spacecraft system. The challenge to future space missions is to identify where machine learning can be applied, which algorithms are most effective, and how to embed these algorithms within the corresponding systems.

The application of machine learning to large data sets is referred to as data mining. Data mining is also referred to as KDD (Knowledge Discovery in Databases). Learning from experience is an inductive process. Data mining is therefore often an inductive learning process – inductive rule learning is common among data mining algorithms, including decision tree rule induction, association rule learning, or case-based rule induction. In all of these cases, a space mission data system may learn (from the data stream) the status of the system, the features of the environment, the scientific discoveries of an experiment, and the anomalies of the system.

Machine learning algorithms may be broadly divided into two categories: unsupervised and supervised. In supervised learning, the system uses training data (with known outcomes or classes) to assign outcomes (decisions, or classes) in response to new incoming data. This often involves the application of a model – a model of the system behavior as indicated by data values. In unsupervised learning, the system learns the rules, classes, features, anomalies, and outliers from the data stream itself, without any preconceived bias or model prejudice.

3.1. Supervised learning algorithms

Supervised learning algorithms include neural networks, decision trees, Markov models, Bayesian networks, and most classification algorithms (e.g., K-nearest neighbors, support vector machines).

3.2. Unsupervised learning algorithms

Unsupervised learning algorithms include principal components analysis, link analysis, self-organizing maps, association mining, and most clustering algorithms (e.g., K-means clustering, mixture modeling).

3.3. Example: application to Mars Rover

We provide an example of an autonomous scientific data collection system operating as a science discovery machine in a remote environment with minimal or no human intervention. In this example, the operational behavior of the science data-collection system is data-driven, through machine learning. Machine learning therefore provides decision support to the autonomous mission systems, in addition to providing decision support to human mission managers. The embedded machine learning algorithms are enhanced through data-sharing (from other sensors, spacecraft, databases, and/or models) – e-Science tools enable this data-sharing – via Web Services, registries, distributed data discovery and access, heterogeneous data fusion, distributed (Grid-like) model computations, and semantic data integration.

To instantiate such an example, we enumerate machine learning applications for a Mars Rover. Numerous machine learning algorithms may be embedded within the roving operational *sciencecraft*:

(a) Supervised learning – search for rocks with known mineral compositions (by classifying each rock sample according to a known list of rock types).

(b) Unsupervised learning – discover objectively what types of rocks and minerals are present, without preconceived bias.

(c) Association mining – find the most common associations (co-occurrences) and also the most unusual co-occurrences of different minerals within rock samples.

(d) Clustering – find the complete set of unique classes of rocks.

(e) Classification – assign rock samples to known classes.

(f) Deviation/outlier detection – find one-of-a-kind, interesting, or anomalous rock/mineral samples.

(g) Learn as the rover goes from sample to sample – build up a model of the environment through Bayesian Networks or Markov modeling. Including spatial tools (such as GIS = Geographic Information Systems) to track the location of samples would provide still greater scientific insight and decision support capabilities.

(h) Information retrieval and fusion – relate the scientific instrument measurement results to other factors, such as dust storms, using data from other *sciencecraft* (e.g., from another rover, or from an orbiting satellite "mother ship").

(i) Decision trees and case-based reasoning – provide on-board intelligent data understanding and decision support (e.g., "stay here and do more" versus "move on to another rock;" or "send results to Earth immediately" versus "send results later").

(j) Case-based reasoning or logistic regression – predict where to go in order to find interesting rocks.

In all of these cases, decisions are based on the incoming data stream, prior experience, new knowledge, and decision logic. The rover can be allowed to act autonomously, without human

intervention, in the deep space environment. Actions are determined by mining actionable data from all sensors. To maximize the decision-making accuracy and effectiveness, the rover should take advantage of other resources. These other resources may include measurements from other data-collecting sensors and models. The latter may be models of the environment (e.g., the geologic origins of the terrain, or the anticipated effects of an impending dust storm), or models of the objects within the environment (e.g., location-dependent rock mineral classes), or models of the *sciencecraft's* behavior. These models can be updated in real-time as new data are acquired – this is data assimilation.

4. Data assimilation & real-time modeling

As new data are collected and integrated with other data streams, the space mission generates and/or updates models of its environment and of its operating scenarios. Updated models (through this kind of data assimilation) can be generated in real-time, thus enabling decision support within the spacecraft itself. The key is to embed machine learning algorithms that learn from the instrument's own collected data stream.

A simple example is a Bayesian Network. The predicted outcome (e.g., decision) of an input stimulus (data stream) is updated as the prior probabilities are updated with new data. The system learns as it goes, through incremental learning. This is real-time modeling of the decision process. If bad decisions are made, the system learns this also. These outcomes can be embedded in a rule-based model, which is invoked through case-based reasoning or a decision tree. The rules of this decision logic represent a model of the spacecraft's science operations – thus, the model becomes more and more accurate through continued data acquisition, machine learning, and real-time data assimilation.

5. Knowledge representation & sharing

Through the acquisition of new data and the development of data-driven models, a space mission can be seen as a knowledge-building system. The system is building new knowledge dynamically about the spacecraft environment and its scientific results. Ideally, the acquisition of this new knowledge is not an academic exercise. Rather, this new knowledge can be shared with other systems, including other space missions or sensors. The computer science research community is actively developing ontologies to represent knowledge and tools (such as the Web Ontology Language, OWL) to share this knowledge. These technologies can be effectively utilized in future space missions to enable heterogeneous space missions to communicate and share knowledge that may help improve cross-mission operations and maximize scientific return.

6. Dynamic data-driven decision support

NASA's space science missions are remarkably productive engines of scientific discovery. The hallmarks of these missions are the depth of science planning in the earliest stages of mission development, the scope of exploration during the deployment and operations phases, and the breadth of new knowledge derived from the science data products.

The challenges of data analysis and exploration are growing as missions become increasingly complex in their instrumentation and as the missions produce exponentially more data in their telemetry packets, engineering streams, and science data systems.

It is an inherited problem with space missions that the computer technologies deployed on-board are necessarily many generations behind what would otherwise be considered leading edge. The timescale to test, radiation-harden, test again, and ultimately deploy on-orbit new hardware (particularly computer hardware) can be one or more decades.

It is nevertheless imperative that technologies that promise increased efficiency and effectiveness of space operations, especially data operations in this discussion, be explored and analyzed as they become available. The hope is that the lessons learned and best practices can then find application in future space missions in order to bring about even greater scientific return. It is in fact particularly useful and relevant to research the application of these techniques to the ever-growing ground-based science data archives, in order to place greater versatility, efficiency, and power into scientists' hands and thus to enable far greater research capability and research potential now, not later. With well documented results, robust proof-of-concept implementations, and unequivocal successes, these technologies may eventually be applied in space, for the benefit of all.

New innovations and growth in the information technology field are profound and rapid. Researching and applying such innovative information systems technologies to the NASA science enterprises can thus pose unique challenges. The technology must be sufficiently well tested and developed that the investment of NASA resources is justified, but it must not be so thoroughly proven that the technology will

naturally find its way into space systems of the future without much investment of NASA's precious resources today. Identifying productive lines of research in the area of applied information systems thus begins with looking at what is bleeding-edge but then proceeds to explore those aspects that are likely to provide a good return on investment in the space arena.

One such area of research is in the application of dynamic data-driven decision support in data-intensive environments, which we have discussed in this paper. In NASA science missions, the volume of data to be processed, analyzed, and explored and the corresponding demands on computational power thus lead naturally to an investigation of the information technology efficiencies that streamline the corresponding compute-intensive and data-intensive operations. Among these technologies are techniques that generate condensed representations of the data stream (e.g., data mining models, as learned through machine learning algorithms) and knowledge extraction from data streams. The extraction of knowledge from the information content of a massive data collection is a clear example of data reduction. This "reduced" knowledge can be communicated and shared among instruments and missions in a much more bandwidth-friendly manner.

Future in-orbit or deep-space data streams may be so voluminous that the data flow cannot be handled. Thus, the abstraction of data-intensive operations (through data mining models and knowledge ontologies) may be the natural solution to the data-volume problem. For example, flight telemetry data streams may be processed and mined for glitches, anomalies, and/or trajectory deviations, and could thus provide the necessary feedback to an intelligent systems loop that corrects the trajectory or other satellite systems appropriately in real-time without human intervention. Similarly, flight engineering data streams can be mined for instrumentation problems or other hardware anomalies and, if possible, yield feedback to an on-board autonomous correction loop. If this correction loop occurs within the computing system, then this is referred to as autonomic computing. And finally, science data streams are projected to grow exponentially in volume, but not all of these data need to be broadcast back to the ground, if appropriate distillations or summarizations are sufficient, and it may also be desired that on-board autonomous data mining and analysis systems perform preliminary processing steps and thus provide feedback to the science planning system – e.g., continue this observation, or stop and look elsewhere, or stop and send results to another spacecraft (such as a member of a constellation spacecraft system). In all of these cases, the data processing, data mining, information retrieval, and knowledge discovery processes are dynamic data-driven decision processes. Applying modern information technology solutions to these processes could have profound positive benefits for future space missions.

7. Some recommendations

In the context of the data-centric technologies described in this paper, we offer a few recommendations for application in future space missions:

(a) Design the measurement system to have a rapid dynamic response capability to incoming data.

(b) Embed data-centric models (e.g., data mining, data assimilation) within the onboard data pre-processing and commanding system.

(c) Apply both unsupervised and supervised learning algorithms within the mission decision support system, and design these algorithms to operate on massive data streams in real-time.

(d) Anticipate that future space missions will be cooperative (e.g., constellations) and will want to share results. Consequently, make the effort to describe data and information content in a higher-level knowledge representation (ontology) format, so as to facilitate data fusion and semantic integration across multiple sensors.

(e) Build data mining algorithms naturally into the onboard processing system: to classify the data; to score the significance of the incoming data; to detect anomalies and outliers; to generate probabilistic alerts from these outlier events; to utilize "learn as you go" models (e.g., Bayes nets or Markov models); and to generate models of the data (e.g., principal components, or descriptive rules, or clusters/classes).

(f) Design the data system from the outset with three modalities of dynamic data-driven behavior: knowledge extraction, decision support, and autonomous systems control.

8. Concluding remarks

The acquisition of scientific data in all disciplines is now accelerating and causing a nearly insurmountable data avalanche. Assimilating these data into models and using these data and models to drive scientific measurement systems are major scientific challenges for today's large scientific research projects. The application of an e-Science paradigm, including Grid computing [6], Web Services, Semantic knowledge representation, and machine learning algorithms will

enhance the scientific return and knowledge-building capabilities of future space missions. These information technologies will enable the science missions to address data-intensive problems that would not otherwise be manageable. This will permit large scientific projects to make use of larger data volumes in the discovery and modeling process than is currently possible. The scientific return on the investments to build, launch, and operate future complex space missions will thus be maximized.

9. References

[1] Eastman, T., Borne, K., Green, J., Grayzeck, E., McGuire, R., & Sawyer, D., "eScience and Archiving for Space Science," *Data Science Journal*, vol. 4, pp. 67-76, 2005.

[2] McDowell, J., "Downloading the Sky," *IEEE Spectrum*, vol. 41, p. 35, 2004.

[3] Plante, R., Greene, G., Hanisch, R., McGlynn, T., O'Mullane, W., & Williamson, R., "Resource Registries for the Virtual Observatory," in ASP Conf. Ser., Vol. 314 *Astronomical Data Analysis Software and Systems XIII*, eds. F. Ochsenbein, M. Allen, & D. Egret (San Francisco: ASP), p. 585, 2003.

[4] Borne, K. D., "Science User Scenarios for a VO Design Reference Mission: Science Requirements for Data Mining," in the proceedings of the Caltech conference *Virtual Observatories of the Future*, p.333, 2001.

[5] Borne, K. D., "Data Mining in Astronomical Databases," *in the proceedings of the conference Mining* the Sky, p.671, 2001.

[6] Borne, K. D., "Grid-Enabled Science with the National Virtual Observatory," in the *NASA Workshop on Grid Computing* (June 2005), downloaded from http://romulus.gsfc.nasa.gov/msst/gridws/ on April 3, 2006.

0-7695-2644-6/06 $20.00