Large Scale Data Mining to Improve Usability of Data – an Intelligent Archive Testbed

Hampapuram Ramapriyan[*a], David Isaac[b], Wenli Yang[c], Steve Morse[d]
[a]NASA Goddard Space Flight Center, Greenbelt, MD 20771
[b]Business Performance Systems, Falls Church, VA
[c]George Mason University, Greenbelt, MD 20771
[d]SoSAcorp, Chantilly, VA 20151

Abstract—Research in certain scientific disciplines—including Earth science, particle physics, and astrophysics—continually faces the challenge that the volume of data needed to perform valid scientific research can at times overwhelm even a sizable research community. The desire to improve utilization of this data gave rise to the Intelligent Archives project, which seeks to make data archives active participants in a knowledge building system capable of discovering events or patterns that represent new information or knowledge.

Data mining can automatically discover patterns and events, but it is generally viewed as unsuited for large-scale use in disciplines like Earth science that routinely involve very high data volumes. Dozens of research projects have shown promising uses of data mining in Earth science, but all of these are based on experiments with data subsets of a few gigabytes or less, rather than the terabytes or petabytes typically encountered in operational systems. To bridge this gap, the Intelligent Archives project is establishing a testbed with the goal of demonstrating the use of data mining techniques in an operationally-relevant environment. This paper discusses the goals of the testbed and the design choices surrounding critical issues that arose during testbed implementation.

I. Introduction

During the last decade, with the addition of the suite of satellites from NASA's Earth Observing System (EOS) Program to the previously existing modes of spaceborne and other observations, a "data rich" environment has been created for the Earth science research and applications communities. For example, by the end of March 2005, the Distributed Active Archive Centers (DAACs) of NASA's Earth Observing System Data and Information System (EOSDIS) held over 3.9 petabytes of data and derived products containing geophysical parameters. These included over 2100 data product types and 60 million distinct files. The data and derived (digital) products were accumulating at the rate of about 2.9 terabytes per day. 2.5 million users accessed DAACs in the year ending in May 2004. The data distribution rate was over 2 terabytes per day.

To realize the full potential of the growing archives of these valuable scientific data, further progress is necessary in the transformation of data into information, and information into knowledge that can be used in specific applications. Such progress is especially necessary given the projected highly distributed capabilities - that includes sensor webs, distributed processing and archiving environments, and distributed communities of users.

Recently, we have conducted a study of concepts for an Intelligent Archive in the context of a Knowledge Building System (IA-KBS) [1] that facilitates the transformation of data into knowledge in a distributed environment indicated above. One of the important aspects of this is a set of intelligent data understanding (IDU) algorithms for data mining and knowledge discovery. Such algorithms are generally viewed as unsuited for large-scale use disciplines like Earth science that involve very high data volumes. There have been many research projects that have developed IDU algorithms with promising results, but they have been tested on data subsets of only a few gigabytes while large-scale datasets tend to be multiple terabytes in size. This paper addresses the design of a testbed to bridge the gap between "research testing" and larger-scale testing of such algorithms to lead eventually to an IA-KBS.

II. Testbed Goals and Research Scenario

Our research on the role and function of an Intelligent Archive within a Knowledge Building System (IA-KBS) began by identifying six key capabilities such a system should exhibit. *Virtual Products* allows a user to treat a product as though it were being retrieved from the archive when, in reality, the data inputs are automatically retrieved, assembled, and processed into the desired form "on the fly," in response to the request [2]. *Event Detection* is used to identify phenomena of interest within a potentially very large data set. Near-real-time reporting of geophysical events in an ingest data stream is one application; content-based queries (e.g., to support the construction of a training set) is another. *Automated Data Quality Assessment* maintains the algorithmic processing pedigree of a data product, and ensures the scientific and algorithmic consistency of the underlying modeling and processing assumptions [3]. *Large Scale Data Mining* is a key component of a knowledge building system and, as will be seen, is a major focus of the IA-KBS testbed. The ability to stochastically optimize the allocation of the storage, network, and computing resources of the archive using *Dynamic Feed-back* supports the other archive functions by increasing throughput and reducing user-experienced latency in product delivery [4]. Finally, the IA-KBS should act as an *Intelligent Requestor* of data, exploiting

[*] This work was performed by the first author as part of his duties as a U.S. Government employee. It was supported by NASA's Intelligent Systems Project and the Earth Science Data and Information System Project. The remaining authors worked as subcontractors under Cooperative Agreement NCC5-645 between NASA and George Mason University. The opinions expressed are those of the authors and do not necessarily reflect the official position of NASA.

its knowledge of information interrelationships and computing resources to minimize its load on cooperating systems.

Based on this six-fold taxonomy of the functional space in which the IA-KBS resides, a wide variety of NASA-funded research projects and their algorithms were surveyed and assessed for their applicability to the testbed. A *reference architecture* for the IA-KBS was prepared in which overlapping or redundant functionality was eliminated, key interfaces and dependencies were specified, and exemplar use cases and associated concepts of operation created and described [5]. The reference architecture provides a way to see clearly where a given research algorithm or capability might reside. The goal was to find research which clearly exhibits both *scientific* and *operational* relevance, and that is applicable to as large a subset of the IA-KBS functional capability as possible. Other evaluation criteria, constraints and considerations that entered into our selection process included implementation feasibility, source data availability, and collaboration potential.

The research effort we selected for the testbed scores very high in almost every area discussed above. The problem selected for demonstration is *fire prediction* [6]. Fire prediction has very high social utility; it is also a notoriously difficult problem, since there is a large and inescapable component of stochastic uncertainty. Fuel type and availability, moisture, sources of ignition, temperature and precipitation – and all these factors measured over considerable lengths of time and seasonal conditions – can affect the solution. Analysis shows that the algorithm touches (either on input or output) most of the functional areas identified in the IA-KBS reference architecture. Finally, the scientific goals of the research will benefit directly from the testbed, since we will be able to process a large variety of geographic areas, fuel types, and seasonal conditions, and hence significantly extend the scientific relevance of the algorithm into new and previously unexplored aspects of the underlying phenomena.

III. Design Issues

The IA project previously identified a number of technical issues associated with implementation of the intelligent archive concepts, most notably scalability issues [6], [8]. The testbed design had to address these issues in a practical way that was operationally relevant and yet could be implemented on a limited budget.

- Scalability and Parallelization. Two approaches to parallelization were considered: coarse grained and fine grained. We selected coarse grained parallelism because, although it requires partitioning the data and generating multiple independent models, it is much easier to run multiple instances of a data mining algorithm on each node than it is to implement a true distributed algorithm. In the fire prediction scenario, we could conveniently partition data along fuel type (land-cover) and month of the year.
- Source Data Restructuring. The fire prediction problem is typical in that data pre-processing is required to put the various parameters into a form and format amenable to data mining, which is very different from the form and format of the source data. Most data mining algorithms, including those considered for the fire prediction problem, consume data in the form of observations, with one record per observation containing all relevant independent and dependent variables. By contrast, remote sensing data is typically stored in files that contain selected parameters for a contiguous region and time. Although an indexing scheme could be used to map from one representation to another, the resulting physical data access would "bounce" around the source data, reducing cache hits and degrading performance. Further, many data mining algorithms iterate over the source data anywhere from a few times to a few thousand times, which would magnify this performance issue. As a result, we chose to physically re-order the data into the form and format expected by the data mining algorithms.
- Representation of Time. The fire prediction problem explicitly involves time as a parameter because of the significant effect of prior precipitation on fire potential. However, the algorithms selected for the testbed have no explicit mechanism for handling time. Instead, observations for the same variable at different times are simply transposed into multiple variables within a single record for input into the data mining algorithm. This temporal "flattening" is straightforward as long as the time series is not too long, in which case the dimensionality of the data space would grow too large. For the fire prediction scenario, we kept the time series short by compressing measurements into a few averages for the prior day, week, month, etc.

IV. Testbed Design

Prior work in the IA-KBS project identified general capability needs, challenges, and opportunities. The testbed design provides an opportunity to explore these general concepts at a practical level that would be relevant to an operational system. The following sections discuss the testbed design. Three different views of the design are provided: a system network view, a functional view, and a software component view.

A. System Network View

The testbed provides a place to demonstrate intelligent archive concepts in an operationally-relevant environment without jeopardizing the production operations of an actual operational system. Features included to make the environment "operationally relevant" include a high-performance node for the data mining and event detection components, use of preproduction and production archive nodes for source data, and high-speed networks for node connections.

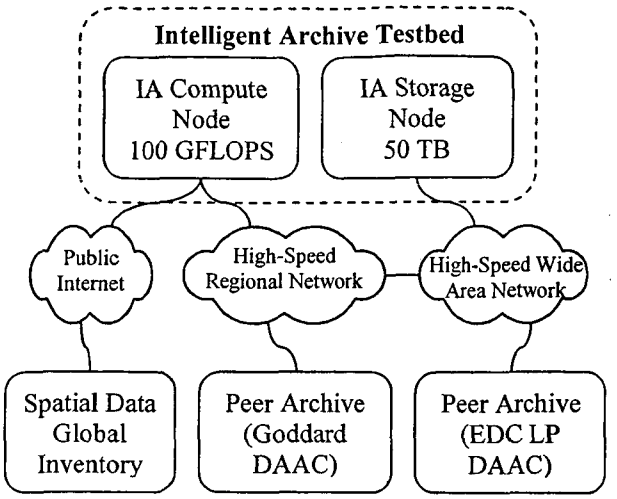

Figure 1. IA Testbed System Components. The testbed leverages existing compute and storage resources to provide enable processing of operationally-relevant data volumes.

B. *Functional View*

The testbed includes a subset of the functional components identified in the IA-KBS reference architecture [5], which were derived directly from the envisioned IA capabilities.

The primary focus of the testbed is on the data mining and event detection components. These two components work as a team: the data mining component examines historical data to extract a statistical model of fire potential; the event detection component then uses this model to scan current data to assess current fire potential.

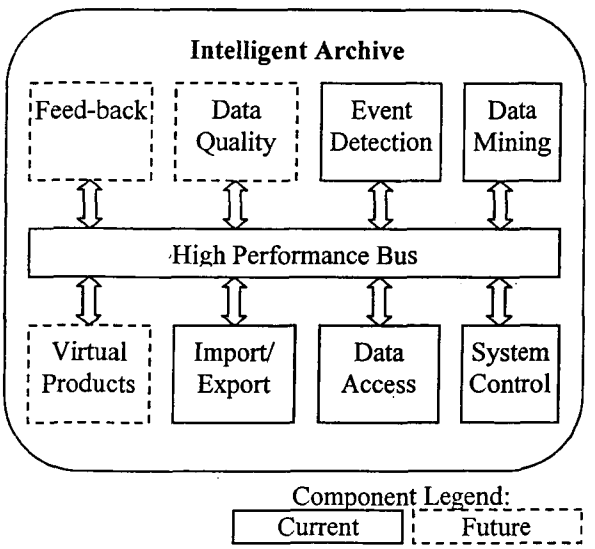

Figure 2. IA Functional Components. The testbed implements most of the IA reference architecture components.

The data mining algorithms process the prepared data and extract a statistical model of fire potential based on the available remote sensing parameters. The first algorithm implemented will be logistic regression, because it is known to perform well in terms of both accuracy and computational effort. Additional algorithms will be implemented as time allows. The data mining algorithms are implemented in MATLAB.

C. *Software Component View*

The testbed includes a variety of software components that together provide the infrastructure needed to manipulate and mine large volumes of remote sensing data.

The Local Application Platform Layer provides four main services. Job management includes the MATLAB Distributed Computing Toolbox/Engine for dispatching different parts of the data pre-processing, data mining, and event detection tasks to different nodes of IA Computational Node. Data access services include HDF libraries for reading NASA remote sensing data files, and MATLAB I/O for storing and retrieving pre-processed data.

The Grid Services Layer provides a variety of services for locating and accessing distributed computing and storage resources. The testbed uses these primarily to identify and obtain source data used by the data mining and event detection components. The Collective layer employs the EOS Clearinghouse (ECHO) for identifying specific files that contain the remote sensing parameters for the times and locations of interest. The Resource layer is used primarily to access data from the grid using GridFTP. The Connectivity layer includes services for authenticating the local server to the grid, as well as the low-level communication services.

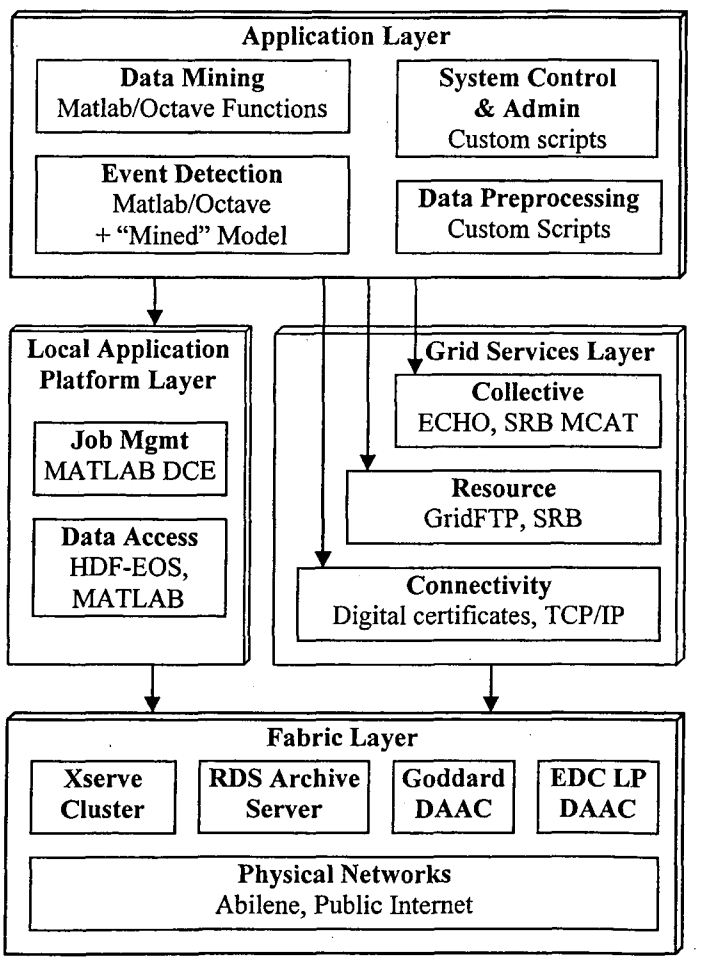

Figure 3. IA Software Components. The testbed provides a suite of services for scalable data mining.

V. DATA PREPARATION AND MINING

As noted above, the first data mining problem selected for demonstration of the IA testbed is wildfire prediction. Fire potential is determined by a number of parameters for which good remote sensing data exist, including fuel type and availability, fuel moisture, current precipitation, and temperature. Ignition events, including lightning and human activities, are the final factor in the occurrence of wildfires, but are excluded from the predictive model because our goal is to predict fires at least several days ahead of time.

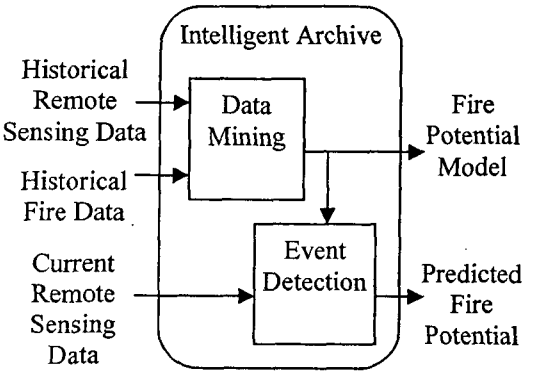

Figure 4. Research scenario. By mining historical remote sensing data related to fire occurrences, we hope to induce a model of fire potential that can be used to predict wildfires.

Initial results have shown that although the induced model has a high error rate, the predictions are much better than chance alone and potentially useful for managing the activities and resources related to wildfire prevention and suppression. The initial investigation scaled back certain parameters (such as spatial resolution) because of the limited computational resources available. The testbed provides an opportunity to seek better scientific results by removing some of these restrictions while demonstrating the feasibility of performing data mining in a pseudo-operational environment.

The basic components of the testbed have been assembled and porting of the data mining code has just begun. We hope to publish results pertaining to the performance of the system and the induced predictive model within the next few months.

VI. CONCLUSION

Design and development of the IA-KBS testbed brought a practical appreciation for a number of issues associated with large-scale data mining, some of which we anticipated and some of which we did not. These include the following:

- Problem-specific approaches to parallelization can be relatively simple but generic approaches are very complex. We chose to partition the problem by parameter value (fuel type and month of year) for simplicity.
- The vast majority of effort in a data mining project involves selecting an appropriate data mining approach, identifying relevant data, and pre-processing the data into a form suitable for data mining. Applying the data mining algorithm itself requires relatively little effort.
- Although data mining algorithms are generally problem independent, the preprocessing of data and the approach to achieving scalability are very dependent on the specific investigation being performed. This points to the need for a flexible set of utilities that can be rapidly configured and applied to a variety of investigations.

A number of unresolved issues are good candidates for future examination. Most importantly, we note that the data mining algorithms for our demonstration scenario have no semantic awareness of space and time dimensions. An investigation of the performance trade-offs (both runtime and accuracy) involved using data mining algorithms that are semantically aware of time and space is warranted given the importance of both to remote sensing applications.

ACKNOWLEDGMENT

This paper was a result of the work by the IA-KBS study team consisting of the authors and the following individuals: C. Lynnes and G. McConaughy of NASA Goddard Space Flight Center; and L. Di of George Mason University. The authors would like to thank David Danks, Brian Bonnlander and Tianjiao Chu, our collaborators for providing the research and applications scenario and initiating the implementation of the algorithms on the testbed.

REFERENCES

[1] H. Ramapriyan, G. McConaughy, S. Morse and D. Isaac, "Intelligent Systems Technologies to Assist in Utilization of Earth Observation Data," presented at Earth Observing Systems IX, SPIE Meeting, August 2004.

[2] M. Clausen, and C. Lynnes, "Virtual Data Products in an Intelligent Archive," White Paper prepared for the Intelligent Data Understanding program, http://disc.gsfc.nasa.gov/IDA/, July, 2003.

[3] D. Isaac, and C. Lynnes, Automated Data Quality Assessment in the Intelligent Archive, White Paper prepared for the Intelligent Data Understanding program., http://disc.gsfc.nasa.gov/IDA/, January 2003.

[4] S. Morse, D. Isaac, and C. Lynnes, "Optimizing Performance in Intelligent Archives," White Paper prepared for the Intelligent Data Understanding program, http://disc.gsfc.nasa.gov/IDA/, January 2003.

[5] S. Morse, and W. Yang, A Conceptual Specimen Architecture for an Intelligent Archive in a Knowledge Building System, White Paper prepared for the Intelligent Data Understanding program, http://disc.gsfc.nasa.gov/IDA/, October 2004.

[6] D. Danks, Biospheric & Ecological Forecasting Case Study: Wildfire Prediction, NASA IDU/AR Workshop, http://is.arc.nasa.gov/IDU/slides/reports04/IDU_Danks_0402.pdf , Feburuary 2004.

[7] G. McConaughy and K. McDonald, "Moving from Data and Information Systems to Knowledge Building Systems: Issues of Scale and Other Research Challenges," White Paper prepared for the Intelligent Data Understanding program, http://disc.gsfc.nasa.gov/IDA/, September 2003.

[8] D. Isaac and G. McConaughy, "Intelligent Archives in the Context of Knowledge Building Systems: Data Volume Considerations", White Paper prepared for the Intelligent Data Understanding program, http://disc.gsfc.nasa.gov/IDA/, September 2004.