Searching for "Unknown Unknowns"

Vickie S. Parsons
Vickie.S. Parsons@nasa.gov
NASA Langley Research Center

Abstract

The NASA Engineering and Safety Center (NESC) was established to improve safety through engineering excellence within NASA programs and projects. As part of this goal, methods are being investigated to enable the NESC to become proactive in identifying areas that may be precursors to future problems. The goal is to find unknown indicators of future problems, not to duplicate the program-specific trending efforts. The data that is critical for detecting these indicators exist in a plethora of dissimilar non-conformance and other databases (without a common format or taxonomy). In fact, much of the data is unstructured text. However, one common database is not required if the right standards and electronic tools are employed. Electronic data mining is a particularly promising tool for this effort into unsupervised learning of common factors. This work in progress began with a systematic evaluation of available data mining software packages, based on documented decision techniques using weighted criteria. The four packages, which were perceived to have the most promise for NASA applications, are being benchmarked and evaluated by independent contractors. Preliminary recommendations for "best practices" in data mining and trending are provided. Final results and recommendations should be available in the Fall 2005. This critical first step in identifying "unknown unknowns" before they become problems is applicable to any set of engineering or programmatic data.

Introduction

The NESC was established to improve safety through engineering excellence within NASA programs and projects. As part of this goal, methods are being investigated to enable the NESC to become proactive in identifying areas that may be precursors to future problems. The goal is to find unknown indicators of future problems, not to duplicate the program-specific trending efforts. The data that is critical for detecting these indicators exist in a plethora of dissimilar non-conformance and other databases (without a common format or taxonomy). However, one common database is not required if the right standards and electronic tools are employed. Electronic data mining is a particularly promising tool for this effort.

Background

NASA has tasked all programs and projects to perform trending as one method to uncover adverse patterns. The NESC has been tasked with performing independent trending across NASA programs and projects. The NASA culture provides a large degree of autonomy and independence for each individual program or project. As a result, a common database of pertinent information, that should be reviewed to identify trends, does not exist. NASA is not alone in this predicament. It has been estimated that 80% of all corporate data is unstructured. Therefore, some electronic mechanism to

extract information from diverse data sources is required. Data mining fulfills this requirement.

The literature contains numerous references to data mining with various, often conflicting, definitions. The General Accounting Office (GAO) conducted a survey to determine the extent that data mining was being used or planned within federal agencies (GAO, 2004). The GAO definition for data mining was "the application of database technology and techniques … to uncover hidden patterns and subtle relationships in data and to infer rules that allow for the prediction of future results." However, in their 199 identified data mining efforts (131 actually operational), they included software that would more accurately be classified as management information systems or general database query languages. One example was the military college ability to determine which students have taken a particular class. Ames Research Center (ARC) considers data mining as "algorithms for executing very complex queries on non-main-memory data." That definition implies that the user has enough knowledge to formulate a query which is not the situation in this discovery of precursors to future problems. Goddard Space Flight Center (GSFC) provides a definition for data mining on their website, "Data mining is defined as an information extraction activity whose goal is to discover hidden facts contained in databases." The problem here lies in the fact that not all the relevant NASA data is contained in a database format. Therefore, this paper subscribes to the definition provided by Frawley et al. (1992): "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data."

Data mining is not the only component of an optimal solution to identify precursors to future problems. Data mining is merely the first step. Once the data mining effort discovers something, the subject matter experts are required to determine if the "something" actually constitutes a potential problem. The discovery of similar events in multiple sets of data may not be an indicator of a future problem. In fact, the "blind application of data mining methods (rightly criticized as 'data dredging' in the statistical literature) can be a dangerous activity easily leading to discovery of meaningless patterns" (Fayyad, 1996, p. 1). Only the subject matter experts can determine which discoveries require further attention. However, the use of domain knowledge experts initially can severely limit discovery (Piatetsky-Shapiro, 1991). Therefore, data mining should be the first step in the overall trending process.

Methodology

To reach a recommendation for implementation of particular data mining software, several steps were initiated. Potential candidate data mining software packages, both commercial and federally-developed, were solicited broadly from contacts within and external to NASA. The software candidates displaying the most promise in the provided descriptions, and documented in previous research efforts, are listed in ***Exhibit 1***.

Exhibit 1. Software Packages Under Consideration

Attensity	**Enterprise & Text Miner**	**Perilog**
Autonomy	Insightful Miner	PolyAnalyst
CART	InSpire	SNOWY
ClearResearch	Intelligent Miner	Starlight
Clementine	Inxight	VantagePoint
DIANE	IR Discover	

The statistical sub-team for the NESC recurring anomaly (RA) effort established a common set of criteria against which the various data mining software packages could be judged. This comprehensive list (***Exhibit 2***) resulted from a consensus of

Exhibit 2. Criteria for Ranking Software Packages

Criteria	Subcategories	Description
Self-documenting		The extent to which the software is capable of documenting the sequence of queries/events that lead to a particular conclusion.
Ease of use		Ease of learning how to use software functions, ease of setting up queries, ease of become proficient, and ease of using it directly. Is on-line training available?
Ease of customization		The capability for build-to-fit or to an individual's unique preference.
Expandability		Ability to expand as the system grows. For example, as more and more data is generated, the program must allow for such data. The system must also expand for additional users.
Minimal preliminary work required		How much data formatting is required? Can the system deal directly with unstructured text?
Visualization		Graphical representation of data using charts, plots, diagrams, etc., to convey information.
Maturity/stability		Program heritage is well established and has a user community that is represented by reputable organizations.
Approach	Clusters, associations, classifications, etc.	The program capabilities represent the inputs from the data mining/statistical community.
Algorithm	Bayesian, neural, vector, etc.	Mathematics behind the program are validated solutions.
Linguistics	Thesaurus, stemming, customized, parsing, taxonomies	Program uses more than a statistical word count … use of sentence structure, etc., to determine "meaning" of words to develop associations.
Server platform	Windows, Mac, Linux	
Client platform	Windows, Mac, Linux, Web-based	
Database compatibilities	Access, Oracle, Sybase, etc.	
Types of input formats handled	Excel, Word, pdf, XML, etc.	
Types of output formats available		
Capability to export datasets		Does the software allow datasets to be exported to another software package for further analysis?
Size of input allowed		Is there a limitation on the number of records/files that can be input to the software for a given "query"?
Minimal pre-selection required		Extent to which the user has to determine the patterns desired as part of the query (i.e., key word search vs. the program "discovering" patterns).
Data types		Whether text, categorical, or variable data allowed.

team member experiences in previous software evaluations, as well as extensive searches into common practices. Each person on the statistical sub-team who attended a software demonstration ranked the software package for each criterion on a Likert-type scale of 1 to 5, with 5 indicating that the package best addressed that criterion. A 5-point scale was chosen based on McKelvie's (1978) empirical research determination that 5-point scales were the most reliable. As recommended by Rantilla and Budescu (1999), three independent experts provided weights for each criterion on a scale of 1 to 5, with 5 indicating the most importance for the complex data mining of unstructured textual data required by NASA. In addition to the three data mining experts, an independent member of NESC and another with interest in data mining from the National Transportation Safety Board (NTSB) weighted the criteria. Because this weighting scheme is ordinal, the median of the five weights was used as the final weighting of each criterion. As King and Elder (1998) indicated, evaluations are unavoidably subjective. Therefore, this mixed approach of using one group to rate the packages and another to weight the criteria, provides some balance to the subjectivity.

Establishing criteria and weighting factors independent from ranking the subjects against those criteria is a well-documented decision-making technique with read-world positive results (Kepner and Tregoe, 1981). The sum of the products of these weights and the median ranks were used to order the software packages. The top four software packages were chosen to be benchmarked against a representative sample of real unstructured NASA and NTSB data. In addition, these packages will be run against a standard dataset, developed and tested by NASA Johnson Space Center (JSC) contractors for use with additional software packages. The objective of this benchmarking will be to determine which software identifies "unknown unknowns", finds clusters that have been previously identified by a human review of the data, operates efficiently, and supports the varied NASA data and infrastructure. The overall process is depicted in ***Exhibit 3.***

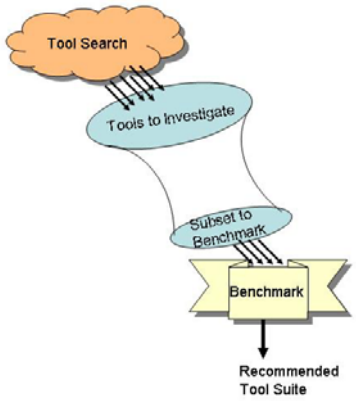

Exhibit 3. Process to Select Data Mining Software

These criteria are clearly not orthogonal because some overlap may be inferred. However, despite the statistical flaws, this evaluation methodology served to reduce inherent bias that would have been present in simple subjective rankings of the software packages by the NESC RA team. A revision of these criteria along with emphasis on cost, vendor support, and prior use satisfaction will be used by the independent contractors chosen to benchmark the top four software packages.

Progress

The demonstrations of data mining software packages are complete. Based on the NESC team's evaluations, the software packages with the most potential are: Autonomy, PolyAnalyst, SAS Text Miner, and VantagePoint. The independent contractors, chosen through competitive procurement, who will perform the benchmarking are Exclusive Ore and Learning Scope. A select group of messy unstructured data from various sources and in different electronic media has been gathered to serve as the data for the benchmarking effort. The contents of this data set are provided in ***Exhibit 4***. Completion of this effort is expected by August 2005. Deliverables will include documentation of benchmark results, final analyses, and recommendations for which software performs best with respect to the NESC needs.

Exhibit 4. Benchmark Data Set

Source	Format	Description
Calipso Project	Text	Problem Reports
GIFTS Project	Excel	Design Review Requests for Action
Shuttle Program	Excel	Problem Reports
Space Station	Text	Problem Reports
Space Station	Word	In Flight Incidents
NTSB	Pdf & Access	Failure Reports
Review Findings	Filemaker Pro	Multiple Project Review Findings
Launches	Excel	Launch Failures

In addition, several pilot studies have been undertaken to evaluate the benefit of data mining in determining "unknown unknowns". NASA Glenn Research Center (GRC) is working with the Safety & Mission Assurance (S&MA) community from NASA Headquarters (HQ) to use Starlight against an occupational hazards database to search for human factors common root causes. NASA Ames Research Center (ARC), as part of the statistical RA sub-team effort, performed data mining, using InSpire, against Shuttle Flight Readiness Review (FRR) presentation material, Shuttle software problem reports, and International Space Station (ISS) data to identify clusters of interest to the subject matter experts. A small set of ISS data was provided to VantagePoint, SAS Text Miner, and ClearResearch, with varied results. Analyses of these results will be compared to the formal benchmarking effort described earlier.

Finally, the GRC Assurance Technology Center (ATC) hosted the second NESC workshop, "Data Mining and Trend Analysis", on March 8-9, 2005, to identify the "best practices" and pitfalls among NASA, academia, and industry in the effort to turn messy unstructured data into valuable information. The results from this workshop can be categorized into cautions, key considerations, and future plans, as shown in ***Exhibit 5***.

Overall, the workshop participants emphasized the need for a consolidated data mining and trending effort within NASA to stop the duplication of efforts which involve a waste of limited resources. The strong recommendation was to create a data infrastructure that is independent of engineering, SMA, and Technical Warrant Holders, but can be used as a resource by any of these groups.

Exhibit 5. Recommendations from NESC Workshop

Cautions	Key Considerations	Future Plans
Start with the end in mind	Potential risks associated with NASA reductions	Create Trending Working Group
Realize there is no "silver bullet" software tool	Role of inadequate training in potential problems	Brief results from software benchmark efforts
Secure buy-in and support from NASA Headquarters	Combination of knowledge management with data mining for maximum results	Brief results from data mining pilot studies
Understand your audience	Use of Failure Modes and Effects Analyses as starting point for cluster identification	Develop continuous improvement strategy
Understand that tool cost may not equate to value	Understanding of data owners	Work to keep community engaged
Present data appropriately – assist in interpretation where necessary	All analyses are not performed by analysts	Create data mining strategy

Plans have been initiated for the NESC to organize a NASA Trending and Data Mining Working Group, with representatives from each of the NASA centers and key experts from academia, industry, and other Government agencies. The purpose of this working group will be to assist NASA in the formulation and implementation of "best practices" for data mining and trending of technical programs and project data, and to ensure appropriate visibility of data mining and trending within the Agency. This will include recommendations for standards, guidelines, tools, metrics, training, and methodologies. In addition, this working group will provide an information resource pool for data mining, trending, and statistical expertise, mentoring and sharing ideas, methods, technologies, processes, tools, and lessons learned to improve communication on trending issues.

Conclusion

A key component in this evaluation of data mining software is the successful application of clustering techniques. Just as Frawley (1992) observed in general, there is a large gap between the generation of NASA data and true interpretation or understanding of the meaning within that data. The data of interest for the NESC independent trending is dynamic, noisy, voluminous and incomplete. In those situations, Frawley (1992) stated that learning algorithms are the most ineffective and discovery algorithms such as clustering are optimal. Advanced statistical techniques alone are not adequate. Berson, et al. (1999) strongly supported the use of clustering as the optimal unsupervised learning technique when the data mining goal is exploration, as is the primary function for this effort. They maintain that the data should define the clusters instead of the user pre-defining areas of interest. Therefore, that capability was a key factor in the software packages selected for further benchmarking. As stated earlier, the discovery through data mining must precede the domain experts' evaluation in order to maximize discovery of those "unknown unknowns."

While this systematic effort to generate the best solution set for the NESC's independent trending task is time consuming, the end results should serve

the Agency across all programs and projects. By simultaneously reviewing data from multiple projects, the NESC will be better able to identify those potential precursors to future problems before they are manifested. In addition, evaluation by the discipline experts only after the software has performed the initial search results is the most efficient and best utilization of the experts' time.

References

Berson, Alex, Stephen Smith, and Kurt Thearling. Building Data Mining Applications for CRM, NY: McGraw-Hill, 1999.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, August 2-4, 1996.

Frawley, William J., Gregory Piatetsky-Shapiro and Christopher J. Matheus, Knowledge Discovery in Databases: An Overview. *AI Magazine*, Fall 1992, p. 57-70.

GAO Report: Data Mining Federal Efforts Cover a Wide Range of Uses, GAO-04-548, May 2004.

Kepner, Charles H. and Tregoe, Benjamin B. The New Rational Manager, Princeton, NJ: Kepner-Tregoe, Inc., 1981.

King, Michel A., John F. Elder IV, Brian Gomolka, Eric Schmidt, Marguerite Summers, and Kevin Toop. A Comparison of Leading Data Mining Tools. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, New York, August 27-31, 1998.

McKelvie, Stuart J., "Graphic rating scales – How many categories?", *British Journal of Psychology*, Vol. 69 (1978), pp. 185-202.

Piatetsky-Shapiro, Gregory, Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop, *AI Magazine*, Winter 1991, pp. 68-70.

Rantilla, Adrian K. & David V. Budescu, "Aggregation of Expert Opinions", *Proceedings of the 32nd Hawaii International Conference on System Sciences*, Hawaii (1999), pp. 1-11.