LA-UR- 98-4077

Approved for public release; distribution is unlimited.

Title: KNOWLEDGE DISCOVERY: EXTRACTING USABLE INFORMATION FROM LARGE AMOUNTS OF DATA

CONF-9809137--

Author(s): Rena Whiteson

Submitted to: INMM/ESARDA Workshop on Science and Modern Technology for Safeguards
Albuquerque, New Mexico
September 21-24, 1998

Los Alamos
NATIONAL LABORATORY

Abstract

The threat of nuclear weapons proliferation is a problem of world wide concern. Safeguards are the key to nuclear nonproliferation and data is the key to safeguards. The safeguards community has access to a huge and steadily growing volume of data. The advantages of this data rich environment are obvious, there is a great deal of information which we can utilize. The challenge is to effectively apply proven and developing technologies to find and extract usable information from that data. That information must then be assessed and evaluated to produce the knowledge needed for crucial decision making. Efficient and effective analysis of safeguards data will depend on our utilizing technologies to interpret the large, heterogeneous data sets that are available from diverse sources. With an order-of-magnitude increase in the amount of data from a wide variety of technical, textual, and historical sources there is a vital need to apply advanced computer technologies to support all-source analysis. There are techniques of data warehousing, data mining, and data analysis that can provide analysts with tools that will expedite their extracting useable information from the huge amounts of data to which they have access. Computerized tools can aid analysts by integrating heterogeneous data, evaluating diverse data streams, automating retrieval of database information, prioritizing inputs, reconciling conflicting data, doing preliminary interpretations, discovering patterns or trends in data, and automating some of the simpler prescreening tasks that are time consuming and tedious. Thus knowledge discovery technologies can provide a foundation of support for the analyst. Rather than spending time sifting through often irrelevant information, analysts could use their specialized skills in a focused, productive fashion. This would allow them to make their analytical judgments with more confidence and spend more of their time doing what they do best.

Introduction

Large amounts of data, from many sources are accumulated in safeguards arenas. These data require organization if they are to be assimilated and understood most effectively by analysts. Unfortunately, today the time-consuming role of assimilating the large amounts of safeguards data still remains largely in the hands of the analyst even though the assessment and analysis of this information should be the analyst's focus. As a result, the importance of automating the assimilation and presentation process is increasing. Moreover, automated processes must not only be able to effectively present the information, but also provide a means of interaction between the information systems and the human analysts.

Proven techniques of data warehousing, data mining, and data and signal analysis, and filtering applied to existing and developing sources of data could provide analysts in every domain area with tools to facilitate their analyses of large amounts of data. Information that is likely to be important according to criteria developed by the analysts can be automatically highlighted and made more readily available, and a ranking of information by likelihood of relevance can reduce the search space that the analyst is obliged to investigate in detail. In limited time, analysts will

DISCLAIMER

anomaly-detection work for the IRS and Medicare, and apply these tools to textual databases, such as those containing published or unpublished papers and reports, or commentary on graphical imagery. Additional tools are being developed and are ready to be applied to data analysis and characterization. By working closely with analysts, computer scientists can customize graphic user interfaces that will address analysts' immediate and longer-term knowledge discovery needs. Throughout the development phase, feedback from the analysts must be regularly solicited to ensure that the technologies that they most need are being developed. Adjustments to an initial plan of work may be necessary to accommodate newly perceived priorities. New interface can be incorporated into existing analyst workstations, greatly enhancing their functionality without intruding into the existing workspace of the analysts.

Background

The United States is currently making a large investment in developing technical data collection and analysis capabilities. Successful utilization of these data depend to a great extent on our success in interpreting large, heterogeneous data sets from diverse sources.

The volume of data collected may be enormous, and the data often vary greatly in reliability and potential relevance. Data— electronic signals and cable traffic, remote sensors, effluent monitoring, and transmission intercepts; open-literature sources; geographic and geologic databases; other measured sources, in the context of prevailing climatic patterns; and human-intelligence source material—may be site specific, but are often general in nature. Increasing data storage capabilities and computer processor speeds have made possible large, integrated databases comprising terabytes of information in a wide variety of formats. As we increase the quantity and complexity of our data feeds, there is an increasing need to support analysts in their effort to assess and analyze these expanding fields of data.

Expertise in the development of analysis techniques, algorithms, and interfaces to provide tools can support our national safeguards efforts . One should also assess the applicability of commercially available technologies, such as keyword and Boolean search engines, as well as technologies that have been shown to be useful, such as analysis by category, *N*-gram analysis, and automated image-quality assessment tools.

In the first stage of knowledge discovery work, it must be determined, together with the analyst community, which proven, appropriate tools could most usefully be applied to existing databases and which areas might warrant the application of new technologies. Every graphic user interface must be customized to analysts' needs through an in-depth interview stage. This interview stage is crucial and provides the detailed roadmap that must be followed to quickly and effectively address analysts' needs. Feedback to and from the analysts must be facilitated, not only at the beginning, but throughout a project, to ensure that what is developed will best serve them. By adapting knowledge discovery technologies to existing stores of data, a reduced assimilation workload for the analyst and a higher probability of isolating and detecting crucial information can be achieved.

Subsequent stages of work involve both expanding information filtering capabilities of the interface tools—for example, providing for the automatic search of relevant databases if an

important visual, electronic, or other signature has been discovered by the system—and application of the interface tools to broader areas and a greater field of potentially relevant data. This might include use of computerized tools to assist the analyst in the decision-making process. Computerized tools can aid analysts by prioritizing inputs; integrating diverse data streams; evaluating multiple conflicting data sources; prioritizing heterogeneous data sources; doing preliminary interpretations; and automating some of the simpler prescreening tasks that are time consuming and tedious. Figure one shows the steps in a typical decision and analysis support system. The analyst's function resides at the top of the pyramid, while the other stages will be integrated into a knowledge discovery system that supports the crucial decision-making function. The selection processes of the system must be transparent, in that the analyst can query system to explain how it is making the filtering or prioritizing choices. The analyst will always be able to override any of the decisions.

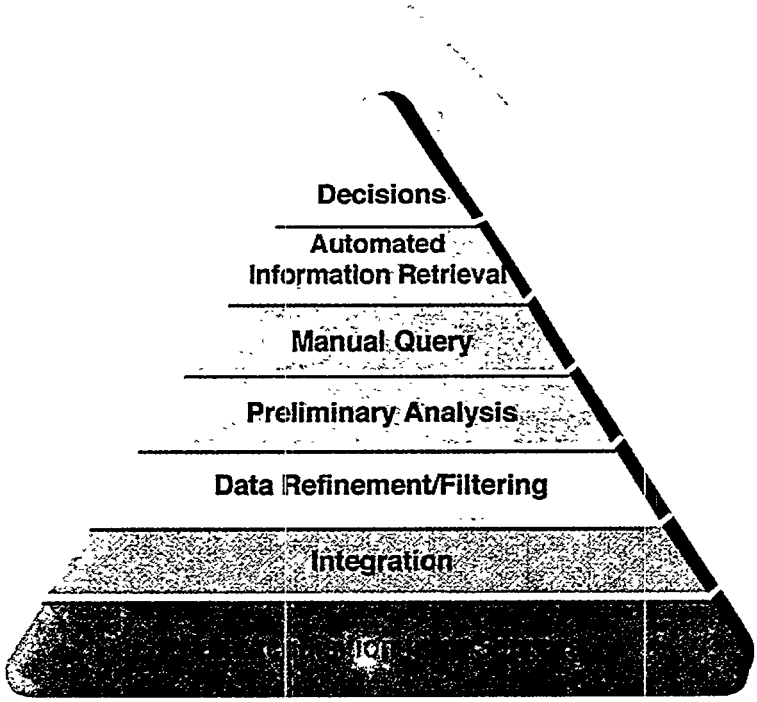

Fig. 1. Steps in a typical decision and analysis support system.

The knowledge discovery technologies provide a foundation of support for the analysts. Rather than spending unnecessary amounts of the finite time available in sifting through often irrelevant information, the analysts could use their specialized skills in a more focused, productive fashion. This would allow them to make their analytical judgments with more confidence and spend more of their time doing what they do best.

The first step in a knowledge discovery project is to adapt existing, proven technologies to the computerized systems used by the analysts as well as the development of technologies that will meet specific needs of the analysts. It has yet to be formally proven that the application of these

technologies will result in greater speed and accuracy on the part of the analysts, but successful application of search and filtering technologies to other large databases strongly suggests that their application to other data sources needs will be useful.

The best approaches will proceed with low risk, well understood improvements while maintaining the ability to bring leading-edge technology to serve analysts' needs where appropriate. First examine the current and planned electronic infrastructure of the analyst community and review the appropriate, available knowledge discovery technologies. Such a review should result in a prioritization of technologies most needed and most readily applicable. It is essential always to maintain the flexibility to adjust efforts if new opportunities or needs become apparent. While providing immediate benefits to the analyst community, there is also the likelihood that synergies will arise between the tools that are introduced, thus producing results that exceed those expected.

Steps to Knowledge Discovery

There are four major methods that contribute to knowledge discovery, though there is some overlap between these methods. These are data warehousing, data mining, information analysis, and automation of expertise. A brief description of each follows.

Data Warehousing enables access to multiple information sources and makes them appear as one virtual database. It thus provides a single interface for all the data. It can also achieve a transformation of a large volume of raw data into a smaller amount with minimal information loss by integrating data from heterogeneous data sources into a single database. While data warehousing is still an area of ongoing research with much interest for academia and industry, it offers several mature and useful components.

Data Mining provides automated methods of detecting properties, patterns, or trends in data. Its methods include:

- Feature Selection: Selecting data fields based upon their relevance or correlation.
- Classification: Categorization of examples into a predefined set of classes.
- Clustering : Discovering the classes of a data set by grouping like examples.
- Pattern Recognition: The identification of forms, shapes, or configurations.

Information Analysis includes computer-assisted analysis techniques that can offer prediction, risk analysis, function modeling, scientific visualization, error location, anomaly detection, trend spotting, finding evidence of non-conformance, and ranking of events.

Automation of Expertise is generally accomplished through machine learning or automated learning. This is done by an algorithm that models a concept or set of concepts by analyzing data. Some of these algorithms can improve their performance based on past actions. These can offer automated parsing of incoming or historical data; preliminary data evaluations; highlighting of key information; and focusing of attention on data of interest. These algorithms include neural networks, which are useful when models and detection methods are not known but require

good training data; expert systems, which require close collaboration with domain experts; genetic algorithms which improve over generations; systems based on fuzzy logic; or decision trees. They can be customized to provide:

- Anomaly Detection: Discerning behavior that is inconsistent with a given model.
- Simulations: Modeling normal behavior and comparing predictions of the model with actual outputs, inventories, etc.
- Statistical Analysis: Starting with mean and variance, the developing models that describe the process under investigation.
- Time Series Analysis: Correlation of data points over time.

Benefits

Potential benefits of a knowledge discovery system include the following:

- Prescreening or filtering of data
- Reduced extraneous workload for analysts;
- Greater clarity and relevance in the information provided to the analyst;
- Alerts for the analyst to intersections of relevant data fields and highlight, prioritize, and provide confidence ranking of preliminary conclusions, and
- Automated retrieval of database information that the analyst has historically indicated has a high likelihood of relevance.

Document Analysis and Data Mining Projects

Safeguards data can constitute a very large, unruly data set not unlike those in databases that Los Alamos scientists have been analyzing in successful efforts to detect fraud, waste, and abuse in Medicare transactions and to detect anomalies and errors in materials control and accountability databases. Document classification methods can categorize documents based on word frequencies or do keyword searches to identify documents in specific subject areas. More sophisticated techniques can apply computational analysis to discover patterns in the data that might otherwise go undetected. The discovery of new patterns in underground facilities data can lead to improved detection and characterization, as it has done in the case of fraudulent Medicare transactions.

With sponsorship from the DOE Declassification Productivity Initiative, the CIA, the American Institute of Physics, and internal R&D funding, Los Alamos scientists have developed advanced methods for document categorization, processing, and retrieval. Given an example document, the query-by-example system can produce from its database other documents with related subject matter. In this approach, the user submits a text of interest and requests that the system find other texts that are similar in content. The system categorizes articles on the basis of previous

examples that have been analyzed and found to have certain common characteristics. This kind of analysis was performed at Los Alamos for Scott and White, a large HMO in Texas, assisting investigators discover clusters of newly emerging diseases or symptoms in the data of patient records.

Los Alamos researchers have applied a query-by-category methodology successfully to the *Physics Review On Line Archives* and to classified documents for the DOE Declassification Productivity Initiative. This technology finds documents that are in the same user-specified category as a model document. The purpose of this technology, in the case of the DOE initiative, is to route documents to the person with the most expertise in that particular area. In ongoing research, the system is being enhanced to rank articles according to the probability that they belong within the various categories of interest.

Query-by-change technology compares images and discovers whether there has been change. Typically this is used when comparing photographs of a site over a period of time. It has been applied to x-ray images of lungs to allow researchers and physicians to visualize the progression of disease. Query-by-anomaly techniques have been successfully used to detect anomalous (often fraudulent) data records in the Medicare program.

Individual analysts may use document retrieval systems in unique ways. If some analysis techniques are more effective than others for certain kinds of searches, it may be useful to understand in what ways they are different. Interesting patterns of analyst behavior are likely to be observable. For the CIA, Los Alamos developed a system that can note these patterns. Similarly, by incorporating feedback on known successes and failures into search systems, it is possible to develop decisions making systems for other data sets that will, over time, become more and more accurate and predictive.

Expert-System Searches for Related Information in Rulebases and Relevant Networked Databases

Expert systems can suggest preliminary evaluations of input data. An expert system is a set of rules that model or codify a process or procedure. Typically, expert systems are designed for specific problem domains. The designer works closely with a domain expert to accurately characterize the process and identify aspects of the problem-solving methods. Los Alamos National Laboratory has developed such rulebases for automated error and anomaly detection and data assessment for nuclear materials control and accountability information systems.[1,2] Developers worked with the domain experts, in this case facility experts, to build the rule base. In an international inspection regime an example rule might be:

[1] R. Whiteson, D. Martinez, B. Hoffbauer, T. Yarbro, C. Baumgart, "Detecting Anomalies in a Materials Control and Accounting Database," *Proceedings of the 37th Annual Meeting of the Institute Of Nuclear Materials Management*, Naples, Florida, August 1996, pp. 955–959.

[4] R. Whiteson, B. Hoffbaur, T. Yarbro, C. Baumgart, "Anomaly and Error Detection in Computerized Materials Control & Accountability Databases," *Proceedings of the 38th Annual Meeting of the Institute Of Nuclear Materials Management,* Phoenix, Arizona, July 1997.

If NO_X is detected

Then Reprocessing, with confidence Y

A set of rules such as this one can be used to automate the process of reviewing incoming data. If a rule is fired, in our example if there is information that NO_X is present, the expert system would report that there are indications of reprocessing at this site along with a measure of confidence in the conclusion. Thus the expert system would be used as a filter to focus the analyst's attention on input data of interest and provide an explanation of the reasoning process that led to the conclusion. The full data set always remains available for direct review by the analyst if desired.

Confidence in a conclusion is a function of the reliability of the source of the data, if available, and the conclusions that can or must be drawn from the data. In this example, the presence of NOx is consistent with reprocessing and raises that possibility, but is not conclusive proof because NOx is produced in a number of processes, such as large power plants, steam driven tools in factories, chemical plants, or in the production of explosives. This information would be encoded into the rulebase and contribute to the confidence calculations.

Existing databases of pertinent technical information that can be networked can be used to provide supplementary expertise for the reasoning system in its making preliminary evaluations about the importance of incoming data.

Independent Components Analysis for Remote Monitoring and Sensor Signal Analysis

There are many developing technologies can be integrated into the knowledge discovery interface once those technologies are mature and have demonstrated their usefulness; one example of such a technology is independent components analysis, wherein the component parts of complex signals are broken out and made available for examination. Such a capability—visualizable as the technology that allows one to isolate a particular voice from the complex sounds generated at a social gathering—could be applied at the workstation to complex data feeds that in themselves yield little or no useful information.

Independent components analysis has been successfully applied to many different signals such as speech (in which multiple voices are separated) music (in which different sound sources can be separated), and brainwaves (in which different processes in the brain can be distinguished). To the extent to which we are able to apply independent components analysis to data from process sensors, the results can be used to identify and characterize the processes taking place.

Challenges

The challenges inherent in development of a knowledge discovery project include the following:

- Adapting known and proven technologies from other problem domains; this must be done in close collaboration with experienced analysts to ensure that these technologies are adapted correctly for the new domain;

- Applying those technologies so that they perform as seamlessly, usefully, and invisibly as possible with the networks and systems already in place;
- Allowing for analyst modification and fine tuning of the tools;
- Development, in concert with the analysts, of expert systems to further expand the analyst's ability to engage large amounts of potentially relevant information; and
- Development of additional, currently unforeseen tools or systems that will be desired by the analyst community as the knowledge discovery project proceeds.

Summary

The goal of knowledge discovery is to provide a decision support tool that extracts knowledge and useful information from large amounts of data. It is not intended as decision maker or replacement for the human analyst. The challenges are many, but technologies exist that can be of use to the domestic and international safeguards community. The key steps are:

- Data Warehousing
- Data Mining
- Information Analysis and
- Automation of Expertise

The rewards of knowledge discovery and information analysis are many. They include prediction, risk analysis, function modeling, scientific visualization, error and anomaly detection, trend spotting, pattern recognition, ranking of events, and discovery evidence of non-conformance.

Additional References

Shirley A. Bleasdale, Thomas L. Burr, James C. Scovel, and Richard Strittmatter, "Knowledge Fusion: An Approach to Time Series Model Selection Followed by Pattern Recognition," Los Alamos National Laboratory report LA-13095-MS (March 1996).

Tom Burr, Justin Doak, Jo Ann Howell, Dave Martinez, and Richard Strittmatter, "Knowledge Fusion: Time Series Modeling Followed by Pattern Recognition Applied to Unusual Sections of Background Data," Los Alamos National Laboratory report LA-13075-MS (March 1996).

Tom Burr and Richard Strittmatter, "Knowledge Fusion: Comparison of Fuzzy Curve Smoothers to Statistically Motivated Curve Smoothers," Los Alamos National Laboratory report LA-13076-MS (March 1996).

Charles Bontemp and George Zagalow, "The IBM Data Warehouse Architecture", Communications of the ACM, September 1998, Vol. 41, number 9

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.