MetaTech Consulting, Inc.

White Paper

A Treatment of Data Mining Technologies

Jim Thomas

June 21, 2003

A Treatment of Data Mining Technologies

Executive Summary

The purpose of the present case study is to expose the reader, from a broad perspective, to a functional grouping of technologies referred to as *data mining.* Following a brief introduction that distinguishes the field between generalist and specialists vantage points, a substantial discussion summarizing an established implementation approach is presented. Subsequently, three different applications of data mining are explored in brief. A concise discussion on the maturity of data mining then follows to highlight both the evolutionary path on which the technology is progressing as well as how mature that path is perceived to by industry evaluation. Consideration is then given to issues of security and privacy that are tightly bound to the technology. Conclusions follow a summarization of the author's views on the application of data mining technology to the challenges confronting the Department of Homeland Defense.

Introduction

The term *data mining* has practical meaning to the generalist and specialist alike. To the generalist, the term refers to any number of techniques by which he is able to examine data – generally located in some form of database(s) – to gain awareness of hidden facts, obscure details, or possibly previously unknown relationships existing in varied data sets. To some, a series of searches on the internet satisfies their definition of the term. To the specialist, data mining connotes a far more discrete set of activities including sophisticated algorithms executed against data sets existing in specialized data structures prepared for the specific activity. The objectives of the generalist and specialist are quite similar though the processes and technologies employed differ dramatically.

Implementation Approach

Mining is performed as a project with specific objectives and requirements. The likelihood of success of a mining project, as with any project, is directly related to the rigor of the process employed. Edelstein (2001) suggests seven steps common to successful data mining projects. Brief summarizations of those steps are presented in the following paragraphs.

Define the business problem.

In order to succeed, the project participants must understand their objective as well as the criteria that will be used to measure their accomplishments. Generally, it is an objective of any data mining activity to discover previously unidentified patterns in some set of data. Beyond that understanding, the project must have a clear understanding of what type of patterns are to be searched for as well as the data – in databases, repositories, or other systems – that they will have access to. With this information, the project team can begin to link together available data to form a logical model or web that provides the breadth of understanding to satisfy the objective.

Build data-mining database.

The project team, having composed a logical model that links together data fields from possibly many different data sources, will work to define a physical data model that can be built to house the data. Data mining relies on statistical analytic process algorithms that access tremendous quantities of data. While relational data models have proven effective for Boolean queries of simple and moderate complexity, they are largely ineffective for data mining. Other structures, such as *Star Schemas, Snowflake Schemas,* and *Dimensional Schemas,* provide a more appropriate structure for the mining problems. It is beyond the scope of this paper to detail these data structures. The author recommends those interested in greater detail on the matter to published works on star schemas and snowflakes (Kimball et al., 1998). Sweiger (2002)

describes the use of dimensional schemas appropriate for more specialized forms of data mining to include *clickstream analysis.* These mining structures have limited application is typical transaction processing environments where relational models are so effective based on their highly de-normalized nature. *Normalization,* the design characteristic of a database to have only a single occurrence of data, is a desirable quality in a relational database that results in efficiencies in storage quantity. Data mining activities are less concerned with minimizing the storage capacity and more concerned with interrelationships of discrete datasets.

Explore data.

Through tools and manual manipulation, the mining team and data analysts go though a process of understanding the source data in its native form. Identification of unsatisfied data needs will drive to changes in the project scope or the access limits to available data. In some instances, it may be necessary to collect new data to build bridges between existing data.

Prepare data for modeling.

After defining the physical schema that will accommodate data from identified repositories as well as satisfy the business problem at hand, the mining team must then capture the process for ingesting the source data. While some data can be brought in as a simple copy, other data must be translated from one form to another or transformed in some other manner. For instance, data from one source may have names of states captured in the long name (e.g. Maryland), others may contain only the two letter digraph (e.g. MD), and still others may provide a zip code (e.g. 20640). To accomplish effective mining, all data must be represented consistently. This transformation is often the most complex of the mining activity.

Build model.

Once the mining database is fully populated with data, data analysts and engineers work together to configure and optimize tools to discern patterns in the data. This is the true mining activity. The algorithms applied use a variety of approaches to discern patterns to include, among others, decision trees, linear regression, neural networks, Bayesian or sequential logic. A treatment of the algorithm strategies is beyond the scope of this paper though detailed works are widely available for consideration.

Building the model is an iterative process that may necessitate changes to the underlying model and changes to the datasets included. The result of the activity is a resulting dataset that represents a model of the source dataset.

Evaluate model.

The model resulting from the application of the data mining tools, with optimized algorithms, is frequently visualized with specialized tools that may be integrated with the mining tools. Through visualization, the analysts are able to interpret the effectiveness of the algorithms and can determine if further iterations are necessary.

Act on the results.

Mining on data is of value only if the results can benefit the organization. The results may be applied to improve processes, better target customers or business partners, or contour the organization's spending plan. It may not be possible to determine just how the results of an effective mining activity might effect an organization until after it has been completed. It is the assertion of this author that, even if an organization undertakes a mining activity for a specific purpose, they are likely to find the results will reveal other, unintended opportunities for improvement.

Technology Application and Use

Data mining is a technology that can be applied to a variety of business problems in many different domains. This author asserts that data mining can be effectively applied in any data-intensive environment. This notion is not novel or unique. Follow are summarizations of some applications of data mining. The initial case provides a generalization of the application of data mining within the *e-Business* domain. The second case is a treatment of the *Glass Box Analysis* environment that employs data mining technologies to address the problems facing analysts in the intelligence arena. The final case briefly describes what is arguably the most ambitious data mining activity yet conceived and, therefore, possible the most controversial – *Terrorism Information Awareness*.

e-Business.

The evolution of e-Business has provided an ideal environment for data mining to continue to mature. In e-Business, data mining has a plethora of business problems as well as an abundance of closely related data. *Clickstream analysis*, a specialization of *Business intelligence*, is explained by Knox and Buytendijk (2001) as "the tracking and analysis of visits to Web sites." This class of data mining has demonstrated the ability to produce insightful profiles of those browsing an organizations web site by generating accurate models of the patterns of their browsing trends behavior.

The application of clickstream analysis to understand the online customers has been described by Ramachandran, Turankhia, and Sripad (2002) by what they refer to as *e-Customer Analytics.* They assert that the information captured on a single online transaction is of more value than the sale itself. This assertion is based on the implicit value of that information to allow the business to more accurately target customers.

In their article on the subject, Imhoff and Norris-Montanari (2000) identify the wide variety of data available to the clickstream analyst and characterized it as follows:

1. Tracking information. This includes the user's identification consisting of a client IP address or proxy server identification. The use of a proxy generally abstracts the user's identity and thus provides less value to the data miner. The customer or user identity, authorized user element used when a secure log-on is required is also available. The request in the form of "GET" or "POST" and the date and time that the Web server responded to the request are also considered tracing information.
2. Server request information about the universal resource locator (URL). This category includes the status of request together with the number of bytes sent.
3. Prior site information. Included in this category are: (a) the URL, (b) the host name, (c) the path, (d) the documents requested, and lastly (e) the query string.

Additionally, details of the customer's browser and client operating system are also available. Collectively, this information provides a very insightful profile of the customer.

Glass Box Analysis.

The Advanced Research and Development Activity (ARDA) for Information Technology has initiated an activity to explore ways to build a knowledge base of what both intelligence analysts and consumers of national intelligence know. The project is named *Novel Intelligence from Massive Data (NIMD).* An architectural component of NIMD is the *Glass Box Analysis* (GBA) element. This element will capture and record activities that occur during the course of analysis, including the stream of analytic tasking, queries, documents examined, and reports

produced. This encompasses the full range of clickstream data described in the case above and extends it with a rich set of other information to included, if possible, the analyst's hypotheses, assumptions, methods, and biases (Advanced Research and Development Activity [ARDA], 2003). While the intent of the NIMD project in general, and the GBA in particular, far exceeds even the most grandiose claims of data mining, the technology is certainly a contributing enabler.

Terrorism Information Awareness.

The Terrorism Information Awareness (TIA) program is an effort undertaken by the Defense Advanced Research Projects Agency (DARPA) to develop and integrate computer technologies that will mine public and private databases to find patterns and associations suggesting terrorist activity. TIA consists of several individual activities – each exploring an enabling technology that can be applied to the problem set. The *Evidence Extraction and Lind Discovery (EELD)* activity is needed due to the inability of existing commercial data mining technologies to address the intelligence-oriented problem. A joint report to congress (2003) by the Secretary of Defense, Attorney General, and the Director of Central Intelligence asserted that data mining technologies of today are capable only of finding broadly occurring patterns and they are not effective at "following a narrow trail and building connections from initial reports of suspicious activity."

Technology Maturity

As the previous discussion has demonstrated, data mining continues to require a significant level of effort and expense and a dedicate instance of data and systems. Though costly and time consuming, it is founded in relatively mature technologies.

Database management system providers have suggested that data mining can be accomplished effectively on the source data systems without the need to build dedicated systems

and without needing conformed copies of data. Figure 1 represents Strange and Friedman (2003) illustration of the relative maturity of DBMS-enabled data mining as a technology. That work suggests this approach to data mining is from two to five years form productivity. This author remains confident that it is unlikely that DBMS-enabled data mining will be effective in complex data environments requiring extensive data transformation and translation. It is possible that approach will be successful in mining cogent datasets that already exist in single databases or are easily conformed.

Security Issues

Many find it alarming just how much information about their time online is exposed for others to exploit. De Lotto (2001) summarized findings of a survey conducted in 2000 by Business Week and Harris Poll that 98% of polled consumers were not comfortable with their browsing habits or shopping patterns linked with their real identities. The American population is even more sensitive to the potential for their personal information to be aggregated by government agencies – even when they perceive the cause to be desirable. In reaction to concern that the program would compile personal information of citizens, Congress imposed a moratorium on the implementation of data mining under the TIA or any other similar program of the Department of Homeland Security. Congress provided a ninety day period for a joint report by the Secretary of Defense, Attorney General, and the Director of Central Intelligence in which they were to detail the program to determine if it would be effective in the war on terrorism while not violating privacy laws or civil liberties (Coyle, 2003). The required report, dated May 20, 2003, presents three program constrains that are believed to satisfactory address both privacy and civil liberties. Those are:

1. Full compliance with all laws and regulations governing intelligence activities and all laws that protect the privacy and constitutional rights of U.S. persons.
2. Development of, as an integral part of the program, new technologies that will safeguard the privacy of U.S. persons.
3. Use for research and testing of either real intelligence information that the federal government has already legally obtained, or artificial information that does not implicate the privacy interests of U.S. persons.

The Electronic Privacy Information Center (EPIC), a nonprofit public-interest group that believes the TIA program could undermine civil liberties and an individual's freedoms, continues to leverage the Freedom of Information Act as a tool to gain access to documents related to TIA to contribute to awareness of the programs implications.

Application to Homeland Defense

Though the scope of implementation may be constrained due to necessary lamentations based on security, privacy, and civil liberties, data mining most certainly does have a future in the Homeland Defense domain. TIA does represent a view of a capability with the most potential to contribute to a national defense. That view continues to struggle for acceptance and continued support. This author suggests that, in the absence of another 9/11 style attack on the homeland, it is unlikely the American citizens will cede protection of privacy and civil liberties to allow the operational deployment of TIA. That is not to suggest a belief that core capabilities of TIA will not mature. TIA is both a research and an integration activity. It is not the development of the technologies that EPIC and other rights advocates are resisting; it is the integration of those technologies to achieve a specific end. Individually, those technologies can benefit the Department of Homeland Defense.

Conclusions

Though concerns over security, privacy, and civil liberties are certain to continue, so too will data mining technologies continue to mature. Brick and mortar businesses and e-businesses alike will continue to capture vast quantities of data that provide the fertile environment for continued development of data mining technologies. As the nation reconciles the conflict between safety and security of its citizens and the rights of those same citizens to privacy and civil freedoms, data mining technologies remain ready to be applied to our homeland defense challenges.

References

Attorney General, Director of Central Intelligence, & Secretary of Defense (2003, May). Report to congress regarding the Terrorism Information Awareness program. Retrieved June 16, 2003 from http://www.darpa.mil/body/tia/tia_report_page.htm .

Advanced Research and Development Activity (n.d.). Glass Box Analysis. Retrieved from June 16, 2003 from http://www.ic-arda.org/Novel_Intelligence/ .

Coyle, M. (2003, June). Fretting over U.S. data collection critics see a lack of privacy protection. *National Law Journal*, 25 (82), p1.

De Lotto, R. (2001, April). Clickstream: Fine to track customers; best at losing them. *Gartner Research Note*, SPA-13-2994.

Edelstein, H. A. (2001, March), Pan for gold in the clickstream. Retrieved June 16, 2003 from http://www.informationweek.com/828/prmining.htm .

Imhoff, C. & Norris-Montanari, J. (2000, August). WhoAmI.com. Retrieved June 16, 2003 from http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=2541 .

Knox, M. & Buytendijk (2001, April). By definition, web analytics needs explaining. *Gartner Commentary*, COM-13-3114.

Ramachandran, P. M., Turakhia, K., & Sripad, R. (2002, October). E-customer Analytics. Retrieved June 16, 2003 from http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=5841 .

Strange, K. H. & Friedman, T. (2003, May). Hype cycle for data warehousing, 2003. *Gartner Strategic Analysis Report*, R-19-9970.

Sweiger, M. (2002, April). Is clickstream data warehousing dead? Retrieved June 16, 2003 from http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=4949 .

Figure 1

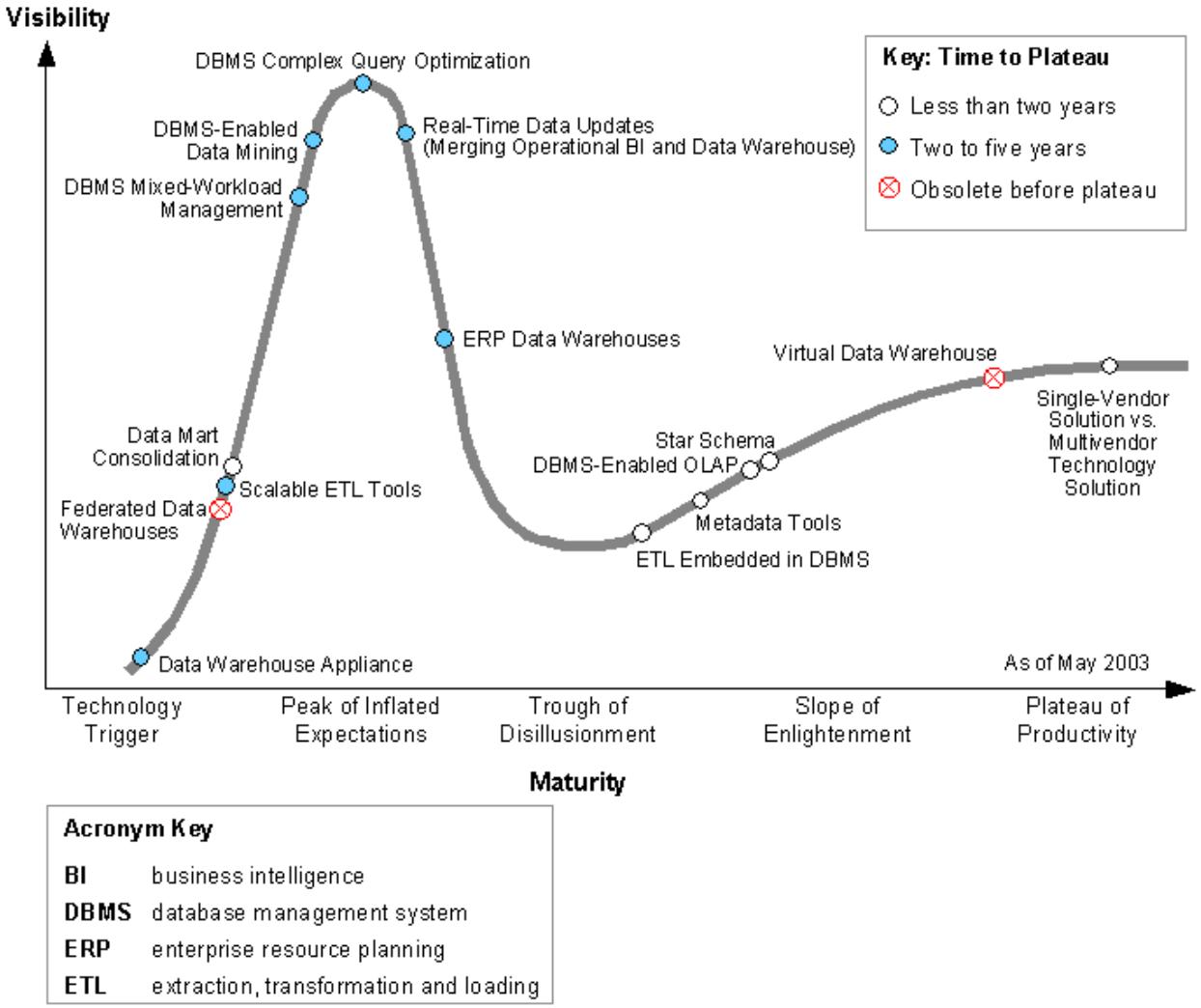

Figure 1. Hype cycle for data warehousing