Data Mining Lessons Learned in the Federal Government

Les Pang
National Defense University, USA

INTRODUCTION

Data mining has been a successful approach for improving the level of business intelligence and knowledge management throughout an organization. This article identifies lessons learned from data mining projects within the federal government including military services. These lessons learned were derived from the following project experiences:

- Defense Medical Logistics Support System Data Warehouse Program
- Department of Defense (DoD) Defense Financial and Accounting Service (DFAS) "Operation Mongoose"
- DoD Computerized Executive Information System (CEIS)
- Department of Homeland Security's Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement (ADVISE) Program
- Department of Transportation (DOT) Executive Reporting Framework System
- Federal Aviation Administration (FAA) Aircraft Accident Data Mining Project
- General Accountability Office (GAO) Data Mining of DoD Purchase and Travel Card Programs
- U.S. Coast Guard Executive Information System
- Veteran Administrations (VA) Demographics System

BACKGROUND

Data mining involves analyzing diverse data sources in order to identify relationships, trends, deviations and other relevant information that would be valuable to an organization. This approach typically examines large single databases or linked databases that are dispersed throughout an organization. Pattern recognition technologies and statistical and mathematical techniques are often used to perform data mining. By utilizing this approach, an organization can gain a new level of corporate knowledge that can be used to address its business requirements.

Many agencies in the federal government have applied a data mining strategy with significant success. This chapter aims to identify the lessons gained as a result of these many data mining implementations within the federal sector. Based on a thorough literature review, these lessons were uncovered and selected by the author as being critical factors which led toward the success of the real-world data mining projects. Also, some of these lessons reflect novel and imaginative practices.

MAIN THRUST

Each lesson learned (indicated in **boldface**) is listed below. Following each practice is a description of illustrative project or projects (indicated in *italics*), which support the lesson learned.

Avoid the Privacy Trap

DoD Computerized Executive Information System: Patients as well as the system developers indicate their concern for protecting the privacy of individuals -- their medical records need safeguards. "Any kind of large database like that where you talk about personal info raises red flags," said Alex Fowler, a spokesman for the Electronic Frontier Foundation. "There are all kinds of questions raised about who accesses that info or protects it and how somebody fixes mistakes" (Hamblen, 1998).

Proper security safeguards need to be implemented to protect the privacy of those in the mined databases. Vigilant measures are needed to ensure that only authorized individuals have the capability of accessing, viewing and analyzing the data. Efforts should also

be made to protect the data through encryption and identity management controls.

Evidence of the public's high concern for privacy was the demise of the Pentagon's $54 million Terrorist Information Awareness (originally, Total Information Awareness) effort -- the program in which government computers were to be used to scan an enormous array of databases for clues and patterns related to criminal or terrorist activity. To the dismay of privacy advocates, many government agencies are still mining numerous databases (General Accounting Office, 2004; Gillmor, 2004). "Data mining can be a useful tool for the government, but safeguards should be put in place to ensure that information is not abused," stated the chief privacy officer for the Department of Homeland Security (Sullivan, 2004). Congressional concerns on privacy are so high that the body is looking at introducing legislation that would require agencies to report to Congress on data mining activities to support homeland security purposes (Miller, 2004). Privacy advocates have also expressed concern over Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement (ADVISE), a data mining research and development program within the Department of Homeland Security (DHS). (Clayton, 2006).

The Center for Democracy and Technology calls for three technical solutions to address privacy: apply anonymization techniques so that data mining analysts can share information with authorities without disclosing the identity of individuals; include authorization requirements into government systems for reviewing data to ensure that only those who need to see the data do, and utilize audit logs to identify and track inappropriate access to data sources.

Steer Clear of the "Guns Drawn" Mentality if Data Mining Unearths a Discovery

DoD Defense Finance & Accounting Service's Operation Mongoose was a program aimed to discover billing errors and fraud through data mining. About 2.5 million financial transactions were searched to locate inaccurate charges. This approach detected data patterns that might indicate improper use. Examples include purchases made on weekends and holidays, entertainment expenses, highly frequent purchases, multiple purchases from a single vendor and other transactions that do not match with the agency's past purchasing patterns. It turned up a cluster of 345 cardholders (out of 400,000) who had made suspicious purchases.

However, the process needs some fine-tuning. As an example, buying golf equipment appeared suspicious until it was learned that a manager of a military recreation center had the authority to buy the equipment. Also, casino-related expense revealed to be a commonplace hotel bill. Nevertheless, the data mining results have shown sufficient potential that data mining will become a standard part of the Department's efforts to curb fraud.

Create a Business Case Based on Case Histories to Justify Costs

FAA Aircraft Accident Data Mining Project involved the Federal Aviation Administration hiring MITRE Corporation to identify approaches it can use to mine volumes of aircraft accident data to detect clues about their causes and how those clues could help avert future crashes (Bloedorn, 2000). One significant data mining finding was that planes with instrument displays that can be viewed without requiring a pilot to look away from the windshield were damaged a smaller amount in runway accidents than planes without this feature.

On the other hand, the government is careful about committing significant funds to data mining projects. "One of the problems is how do you prove that you kept the plane from falling out of the sky," said Trish Carbone, a technology manager at MITRE. It is difficult to justify data mining costs and relate it to benefits (Matthews, 2000).

One way to justify data mining program is to look at past successes in data mining. Historically, fraud detection has been the highest payoff in data mining, but other areas have also benefited from the approach such as in sales and marketing in the private sector. Statistics (dollars recovered) from efforts such as this can be used to support future data mining projects.

Use Data Mining for Supporting Budgetary Requests

Veteran's Administration Demographics System predicts demographic changes based on patterns among its 3.6 million patients as well as data gathered from insurance companies. Data mining enables the VA to provide Congress with much more accurate budget requests. The VA spends approximately $19 billion a