DATA MINING LITE

"We are drowning in information
and starved for knowledge"
John Naisbitt

by
David Chereb
John H. Mercer

DATA MINING LITE

by
David Chereb
John H. Mercer

"We are drowning in information
and starved for knowledge"
John Naisbitt

If you're Bill Gates or Michael Eisner you can afford data mining today. Data mining is beginning to help large businesses find information hidden in existing data. These companies are drowning in data. With millions of customers and billions of transactions every week, they need massive 'data warehouses' and specialized 'data mining' tools.

But, you need to maximize profits as much as they do. Discovering hidden relationships in your data is just as vital for your business. 'Data Mining Lite' is our term for applying data mining techniques for smaller firms.

The quote from John Naisbitt is right on target. Data mining is the light at the end of the data tunnel. This article will show what data mining is, how much it costs, and whether it makes sense for you.

WHAT IS DATA MINING?

Data mining as currently defined: ***"Data Mining" is the process of finding hidden relationships in large databases.***

For instance, consider a mass mailing. Everyone knows that an untargeted mass mailing only gets a 2% (or less) response rate (i.e. I send out 1000 letters and get back 20 replies). If I target the mailing based on data mining discoveries, I can raise the response rate to say 5%. This means 1) I can send out fewer letters and save money, or 2) I will get more sales from each mailing. Either way I win. That's why companies are starting to data mine.

The most important part of the definition is finding the hidden relationships—that is the payoff to all of the effort that goes into a data mining project.

Large Databases

The reference to large databases is unimportant conceptually, but very important during the implementation phases. Data mining tools are currently designed to analyze millions of records, such as customer buying habits. Think of American Express trying to determine what items for sale to stuff into their monthly bill to you.

Data Mining Tools

In order to find these hidden relationships, the data mining tool automatically searches the data for correlations; i.e. a connection between two or more items. For example, the result of a data mining run might be: men between 35-45, who have bought golf clubs in the last four months are also likely to go on two vacations in the next six months.

The main tools used in data mining are: visualization, regression analysis, neural networks, classification and clustering. The table (see Analytical Tools section) defines these tools in more detail. For now the only thing that is important is that these tools find relationships that are valid and can be utilized to increase profits.

WHEN DOES IT WORK?

Just because you use data mining doesn't mean your profits will increase. Think of just two cases where it isn't very helpful; 1) you don't have any data and 2) the results do not find anything interesting that you don't already know.

Unknown Relationships

The definition says that data mining searches and finds 'unknown' relationships. That is true and is the standard way data mining is viewed. But in reality the main benefit of this entire process is benefiting from your data. The more general payoff to data mining is that you will begin to view your data as a strategic resource.

Most of the time the results of data mining are not startling. The reason is simple: if you have been in business for awhile and have been successful, you already know a tremendous amount about what drives your customers. Data mining can find additional relationships, but they will tend to be 'second order' effects. First order effects are the big ones you already know: high income people buy expensive cars and houses; older people take more cruises; women buy more cosmetics, etc., etc. Second order effects are less apparent.

Very Large Databases

Data mining emerged because the amounts of information companies are gathering is skyrocketing. Daily transactions may reach into the tens of millions. Analyzing this much data with conventional statistical tools isn't practical. Data mining uses sampling

and automatic search techniques, to reduce the time and computer horsepower required to run an analysis.

Data mining lite is applying the same techniques to much smaller databases (ones that may be only 1 million records or less). The statistical results of data mining lite are just as valid as data mining 'heavy' procedures.

WHAT DO YOU NEED TO ANALYZE YOUR DATA?

Before you discover interesting things through data mining, you need to meet some minimum requirements. Some of these are:

Data

Data mining needs data. If don't have any data, there is nothing this tool can do for you. It is like a car out of gas. Until you put in the gas, you don't go anywhere. But having the data isn't enough. Not all data is created equal.

Clean Data

In most cases we are just keeping our heads above water with our operational data. The database is an Excel spreadsheet that has gotten bigger and bigger. The file structure was never designed for today's questions. The query process is a series of new reports. The databases are all over the place, and they run on different hardware and software platforms and do not talk to each other.

In this typical scenario, data mining is of little help until the databases are 'cleaned up'. In some cases the majority of the data mining effort is spent cleaning up the data so that it is valid. That is, removing obsolete product codes or duplicate customer addresses or 'ambiguous references' (see sidebar). This process is usually incorporated into an expensive data warehouse. A data warehouse usually helps to organize the data into a format for quick analysis. The first step to minimize the clean-up efforts required is to collect operational data efficiently and store the information in a well designed database.

CHARACTERISTICS OF A GOOD DATABASE

With the proliferation of powerful database software on desktops nationwide, it is becoming easier to build a good relational database. The further away your database is from the listed characteristics, the more time and money it will take to data mine and use the data for effective decision making.

Item	Description	Example
Single Data Source	Each table, which holds information, is the only table in the organization, which holds that data.	The human resources department maintains records with employee names, addresses and position information. Accounting and each department with employees use this information instead of recreating their own databases.
First Normal Form	Each field in a table describes a single characteristic of the item being listed in the table.	For each order, order items are stored in a separate table listing one order item per record. The order item information is stored in the same respective fields. Fields do not repeat horizontally (i.e. Item1, Qty1; Item2, Qty2.)
Second Normal Form	Each record in a table contains a Primary Key Field, which uniquely identifies the particular item being listed in the table. No field exists in the table which does not depend on the primary key.	Each employee is given a unique employee ID number. The table in which this employee ID number is assigned contains fields which only hold information about the employee (i.e. Name, address, SSN.) Department information or job descriptions are not found in the employees table.
Entity Integrity	Each table's fields contain data which only describes the item being listed in the table. The Primary key field must contain a non-null, unique value.	Customer information (i.e. name, address, phone) is stored in one table, and sales information is stored in another table without reentering customer information for each sale. There is a unique value for CustomerID in each customer record, and it is never empty.
Primary Table (lookup table)	The table in which the Primary Key value for each record is generated or assigned. This table holds "Parent" records.	The Employees table is a Primary table. The EmployeeID is the only key field in the table. Employee information cannot be gotten to except through reference to the Primary Key in this table.

Item	Description	Example
Secondary or Related Table (transaction table)	A table which holds, in addition to its own primary key, one or more fields which store a value equivalent to a records primary key value in another table; these are foreign keys. This table holds "Child" records.	The JobAssignments table contains JobAssignmentID, EmployeeID, PositionID, DepartmentID, and StartDate fields. The EmployeeID, PositionID and DepartmentID are foreign keys that allow Employee, Position and Department information to be looked up via their respective tables' primary key fields.
Relationship or Link	A connection which exists between two tables which functions as a link between the Primary Key in one table and the Foreign Key in another table.	
Third Normal Form	Non key fields describe a mutually independent attribute of the item being listed. There are no calculated fields stored in the table. Lookups to other tables are done only by the foreign key field without repeating descriptive information.	An order record contains the price per unit and the quantity of units sold but not the total (quantity x price) for that item. The same record contains a field listing the ItemID, which is the primary key for that Item in the ItemsForSale table.
Referential Integrity	A rule which prohibits a foreign key in a related table from existing without a corresponding Primary Key in a primary table. No "Orphan" records allowed.	For a Social Services Department a client record cannot be added without a caseworker being assigned to the client in the clients table. Nor can a caseworker be deleted from the caseworkers table if any client is assigned to that particular caseworker.

Consistent Definitions

If you haven't dealt with any large databases that have been in use for five or more years, this topic may seem obvious to you. Sales should be sales; a product should be a product—for all company locations and products. So far this sounds like accounting and to some extent it is. Consistent definitions make much of number crunching and accounting tasks, easy.

Making older databases consistent has been a major part of the data mining effort. There is no one reason why so much inconsistency exists; it's just that no one thought all these separate databases (sales, accounting, manufacturing, vendors, etc.) would all be 'talking' to one another. But in today's networked environment, inconsistent data definitions pop up all the time. If you have a well defined database (following the rules described in the prior table) you are about 70% of the way to doing data mining (lite).

Analytical Tools

The specific tool that is used to discover patterns is less important than always remembering that the focus is finding valid, stable, causal patterns that are strong enough to use for enhancing profits. The techniques described next are some of the most popular data mining tools available.

Many times in data mining and data mining lite, more than one tool is used (especially if more than one tool is included in your software package). If the answers re-enforce each other, it gives you more confidence that you can apply the results.

DATA MINING STATISTICAL TECHNIQUES

Method	Characteristic	Advantage	Disadvantage
Regression	Uses statistical tests to show importance of each variable	Shows how important individual variables are.	Difficult to test for all combinations of relationships
The term regression comes from statistics and genetics. In genetics there is a tendency for offspring to 'regress' toward the norm. For instance, two eight feet parents tend to have tall children, but not as tall as they are. Two five feet tall parents, tend to have children that are taller than they are. That is, children tend to regress toward the norm for human beings. If this were not so, we would see at least some fifteen feet and some one foot tall human beings—but we don't. The part of regression that is relevant for data mining is that 'driver' variables are used to 'explain' the variation in some other variable. If I want to explain when home buying goes up in an area, I would probably include such driver variables as: interest rates, job growth, migration patterns and more. The result would be an equation that explains why home buying goes up and down, by how much and which driver variables are important. The advantage of regression analysis is that it is very flexible, is theoretically sound, and it is easy to understand and explain the results.			
Decision Trees	Repeatedly classify characteristics into a few branches.	Quick, easy to follow	Overstates divisions among data
Decision trees separate and classify data points according to their behavior. In many ways this technique is similar to k-nearest neighbor (with the advantage that it takes less number crunching). A decision tree attempting to find who buys rock magazines would first look at age and divide the data into those more likely to buy and those less likely. Next, income might be used. Frequently only two or three levels are needed to reach meaningful results.			

Method	Characteristic	Advantage	Disadvantage
Neural Networks	Mimic brain using many simultaneous connections to search for patterns	No prior knowledge of what to search for is needed	No explanation of how the results were arrived at
Neural nets for data analysis are a recent development and are mimicked on the way human brains are wired. In our brains the neurons are connected to many other neurons through dendrites. The magnitude of the electrical and chemical signal between two neurons is a measure of their closeness. In solving a problem, many neurons fire at the same time and the answer is the sum of the network. Neural nets can frequently be used as a substitute for regression analysis. The advantage of this new technique is that less expertise is required to set up the analysis. One drawback is that explaining the results is more difficult (but applying the results is not difficult). In a neural network the 'equation' looks very similar to a regression equation; i.e. there are many inputs and each one has a weight associated with it. By entering values for the input variables and attaching the weights, an answer is produced			
k nearest neighbor	People with similar characteristics (such as age, income) behave the same	Easy to understand results	Lots of number crunching
'Birds of a feather flock together.' k nearest neighbor uses this to predict behavior. This technique is easy to apply but uses a lot of number crunching. This technique assumes that people who subscribe to the same magazine, are similar. That people who shop at the same store are similar. A typical result might be: young, lower income people tend to read rock magazines, while older, higher income people, read golf magazines.			
Genetic algorithms	Uses 'survival of the fittest' technique to find best correlations. Can usually be used in same situations as Regression & Neural Networks.	Efficient at finding good solution (but it may not be optimal solution)	Requires user to set up parts of search methods & not widely available (early 1998)
Genetic algorithms emulate Darwinian evolution in that the algorithm takes the 'survivors' among competing data arrangements and combines these winners into a new arrangement. This is done many times, until no meaningful gains are produced. The resulting arrangement is used for prediction. This tool searches for the best arrangement of weights to apply to explanatory variables. In some ways it is similar to regression and neural networks, in that it tries to find the best combination of weights to predict behavior.			

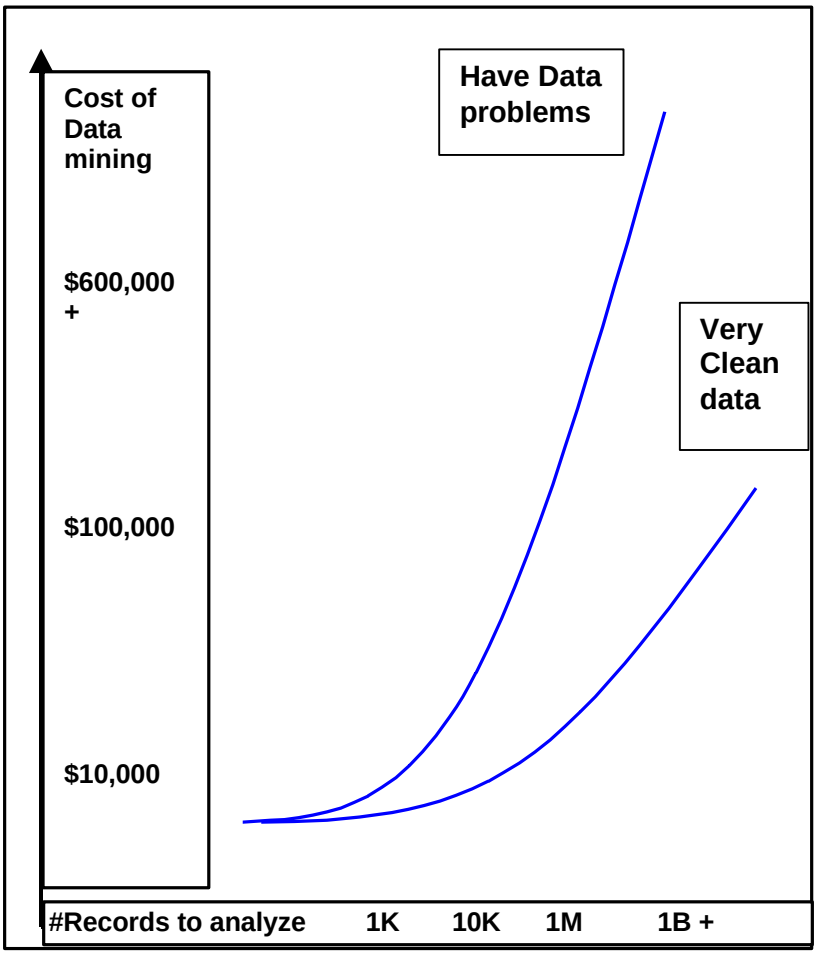
Business Judgment

We want to emphasize that data mining results cannot be blindly implemented. The results must be looked at by seasoned managers who know how the business runs. Often the results will reinforce management's intuitive decisions. Do the results make sense? If not, do not use the data mining results. In most cases, the results will make sense and can be used in some fashion.

HOW DO YOU MAKE MONEY WITH DATA MINING LITE?

Significant monetary gains are possible with data mining lite. The gains generally fall into two categories: 1) Marketing/Customer Service, or 2) Internal Controls.

Improving service increases customer satisfaction and profits. Targeting your products and services to closely match what potential customers want was the initial example of how data mining lite is profitable. Data mining has also discovered unusual patterns in cash sales. These small deviations from the normal receipts indicated fraud, and they had escaped normal accounting controls.



Apply Results

To make money with data mining, you have to implement the results. While this seems trite, many projects suffer from 'paralysis by analysis'. We suggest you do the data mining; carefully examine the results to see if it makes business sense; then do a trial run (if possible) before going all out. If the data mining results have been robust, you will enjoy much success with this tool.

HOW MUCH DOES IT COST?

To date, data mining is an expensive undertaking. Most projects have cost well over $50,000 and can run into millions. It doesn't have to be this way and it is rapidly changing. The key parameters determining how much it costs are the amount of data to be mined and how clean the data is. The next few paragraphs will show whether your own data mining project is likely to cost $10,000 or $100,000+.

More Records, more money

In general if you need to analyze 200 million records, it will be more expensive than if you only have to look at 10,000 records. In the first case, which is a typical data mining scenario, you need fast expensive computers, a great database (probably a Data Warehouse) and specialists (data miners, database administrators, etc.).

If you only have a few hundred or a few thousand records, you can use desktop tools such as MS Access and MS Excel to analyze the data. Many practitioners would say this is not data mining; but so what; the business goal is to find interesting patterns in the data and profit from them. That's why we call it data mining lite.

In addition to internal data there are vast amounts of secondary data, information gathered from outside the firm. Outside data frequently augments internal data in two ways: 1) to fill in the data holes when you have no data (i.e. you can't sort by income because you don't have that information for customers on the company mailing list). The second use of external data is to do a crosscheck on internal data, to check for reasonableness. Using secondary data adds to the cost of data mining, but it also greatly expands the potential payoff.

Clean Data, less money

If you are like most companies, you have lots of data. Most of it is probably financial data with a little sales, production and marketing information thrown in. But it is highly likely that the data is not ready for data mining. Data mining requires 'clean' data. This is not trivial and it is not common to find clean databases.

More Existing Hardware, less money

Today you can perform most data mining projects from the desktop. This was not possible just six months ago. Two changes have made this possible: 1) desktop computers are more powerful; a 300+MHz PC with 64+Mbytes of memory will do just fine as long as it's networked to the source database. 2) Data mining tools are being downsized to work more easily with desktop computers.

Within twelve months an entire data mining lite project, database and analysis, will be done solely with a desktop machine.

Existing Data Warehouse, less money

Cleaning up the existing data is the most time consuming and expensive part of the data mining project. If you already have a large 'data warehouse' with clean, consistent data, you will save 70% of the cost of starting from scratch. Even if you don't have a data warehouse, you may still be OK, as long as your operational data is clean.

There is a difference between arranging data for analysis and tracking it for accounting. Accounting keeps track of each transaction in order to post it to the books, and everything is linear or a simple ratio. However, data for analysis is designed to ferret out relationships among data items; i.e. correlations. These relationships are not linear, such as C=A+B, but rather more like C is impacted by some nonlinear combination of A and B such as log(C)=A*B*X-D/Z. Thus a transaction database keeps track of how many of A, B and C we sell each month. The analytical database keeps track of information that will tell us why sales of B increased or why A and C move in opposite directions most months. The analytical database can be used for realistic sales and profit projections, instead of the linear extrapolations used by accounting systems.

SUMMARY

Data mining is an exciting new business productivity tool. In the right circumstances the results of data mining can yield a high ROI. During the next few years, the cost of data mining will drop as users begin to use 'data mining lite'. User experiences from companies of all sizes will reveal valuable new lessons on the strategic use of data.

In order to use data mining you need clean data and fast data analysis capabilities.
If you choose to look further into data mining, excellent information can be found at several of the web sites listed below.

FURTHER LINKS & REFERENCES

More information about data mining, along with links to many data mining web sites can be found at: DCG, Inc. (the authors' company) or one of the other web sites listed.

http://www.davidcherebgroup.com
http://www.dmreview.com
http://www.data-warehouse.com

The Authors:

714-381-8088

David Chereb, Ph.D. (economics) is recognized for his advanced approach to business dynamics. His concentration is market assessments and international acquisitions. His latest book (available spring '99), "Strategy and Evolution: Profiting from Business Intelligence Tools" expands on the theme of this article to include OLAP and Decision Support Systems.
dc@davidcherebgroup.com

John Mercer, MBA is an expert in database design. He frequently trains analysts on advanced database tools.
johnm@davidcherebgroup.com