Data Mining with Silicon Graphics Technology

Arun Swami
Member of Technical Staff
Silicon Graphics Computer Systems

Abstract

The efficient construction and verification of models of patterns embedded in large databases—data mining—is emerging as a key technology for enterprises that wish to improve the quality of their decision making and gain competitive advantage by exploiting operational and other available data. This paper outlines some concepts and techniques of data mining, contrasts data mining with related technologies, and describes how Silicon Graphics CHALLENGE™ servers are the platforms of choice for data mining.

Introduction

Competitiveness increasingly depends on the quality of decision making, so it is no wonder that companies often try to improve the quality of decisions by learning from past transactions and decisions. To support this process, operational data generated by business transactions is consolidated in a *data warehouse*, which is often a relational database system (RDBMS). Users of the data warehouse should know what questions to ask in order to identify patterns and gather needed information. *Data mining* facilitates further exploration by discovering the implicit patterns in the data. Hence, data mining significantly augments the value of the data warehouse.

Data mining is relevant to many different types of businesses. For example, retail stores obtain profiles of customers and their buying patterns, and supermarkets analyze their sales and the effect of advertising on sales. Such “target marketing” is becoming increasingly important. Data mining has also been applied in industries such as health care, insurance, and finance.

This paper describes data mining and some of the techniques used in data mining, presents example applications, and mentions some of the prerequisites for employing data mining effectively in a business operation. This discussion shows that Silicon Graphics CHALLENGE servers are attractive data mining platforms.

1.0 Technologies Related to Data Mining

To make data mining feasible, the appropriate data has to be collected and stored in a data warehouse, and adequate system resources have to be available to make the data mining process feasible.

Statistical analysis systems such as SAS and SPSS have been used by analysts to detect unusual patterns and explain patterns using statistical models such as linear models. Such systems have their place and will continue to be used. Data mining techniques will not replace such analyses but, in fact, may spur more directed analyses based on the results of data mining.

Ad hoc querying and report generation are commonly used by many businesses to provide input to their decision making. Decision support systems (DSS) and executive information systems (EIS) are often used for this purpose. At a more basic level, report generation tools are used. Applications utilizing the ad hoc querying capabilities of relational database systems are ubiquitous. Data mining helps focus the use of these systems and techniques so that relevant information is obtained faster and the analyst's time is employed more effectively.

Multidimensional spreadsheets and databases are becoming popular for data analyses that require summary views of the data along multiple dimensions. As with the methodologies described above, the analyst has to provide continuing guidance for this process to be successful. Data mining technologies perform automatic analysis that can help enhance the value of the data exploration ("drill down", "drill up", summarization along various dimensions) supported by multidimensional tools.

Neural networks have been applied successfully in a few applications that involve classification (see Section 3.2). However, they suffer because the resulting network is viewed as a black box, and no explanation of the results is provided. The lack of such explanations inhibits confidence, acceptance, and application of results. Also, neural networks suffer from long learning times, which become intolerable when large volumes of data need to be processed.

The enabling power of data mining is even more clearly seen with respect to another nascent technology, *data visualization*, which can make it possible for the analyst to gain a deeper, intuitive understanding of the data. Data mining can enable the analyst to focus attention on important patterns and trends and explore those patterns and trends in depth using visualization techniques. Data mining and data visualization work especially well together. Data visualization by itself runs the danger of being overwhelmed by the sheer volume of the data in commercial databases. Data mining may help by suggesting a starting point for fruitful exploration and the appropriate metaphor to use for the visualization.

Section 3.1 and Section 3.2 present examples of the kinds of discovery that data mining technology makes possible. Data mining, while complemented by the techniques described above, adds significant value beyond the use of the traditional techniques.

2.0 Data Warehouse

Collecting the data to mine is a difficult process by itself. While operational data is often collected for business processes, once the immediate use of the data has passed, the data is often archived. Such data can now be exploited for data mining; however, a suitable system infrastructure requires high I/O bandwidth for fast loading and parallelism to reduce the time for initial loads or updates.

Figure 1 illustrates the data sources and processes leading to the data warehouse. The parallel streams indicate that cost-effective, scalable parallelism is a critical technology for the data warehouse.

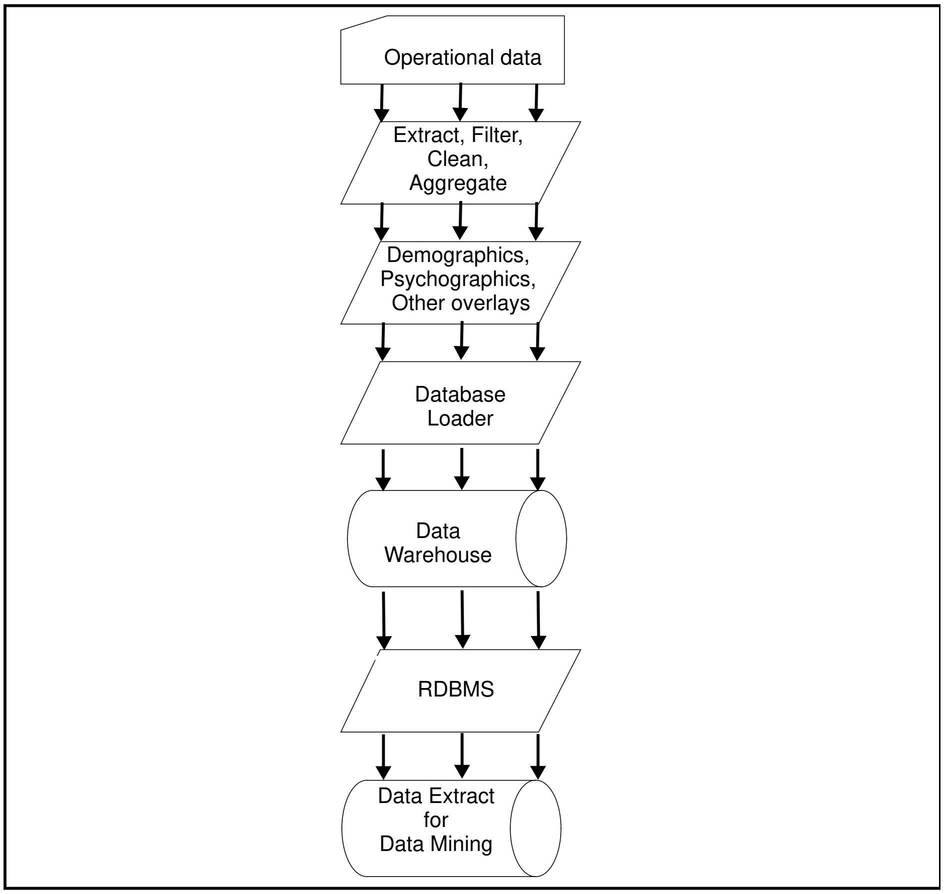

FIGURE 1 Data Warehouse for Data Mining

The data extraction process extracts useful subsets of data for mining. Sampling may limit the size of the extracted data, and selection predicates may be applied to focus on applicable subsets of data. Aggregation may be done if summary statistics are useful.

Operational (or transaction) data is often not in the most suitable form for mining. The data needs to be cleaned, with incorrect fields being corrected or otherwise marked. This cleaning process can be quite sophisticated, with a lot of business knowledge being encoded in the cleaning rules. Fields that are not necessary for the analyses may be omitted in order to reduce storage requirements and speed access for analysis. Some degree of normalization may be performed to reduce the amount of redundant data. Conversely, some denormalization may be performed to speed up data access. Additional fields may be added using "overlays". For example, demographic overlays are commonly used in marketing research. Matching rules are needed for the overlays, and these rules are specific to the business domain.

3.0 Data Mining Examples

Data mining technology has evolved in a bottom-up, application-driven manner. This section describes two important data mining technologies: mining for association rules and profile generation.

3.1 Mining for Association Rules

Association rule technology has many applications, including supermarket inventory planning, shelf planning, and attached mailing in direct marketing. For example, association rules can be derived from mining "market basket" data for buying patterns exhibited by customers. Here, each "market basket" consists of the set of items bought by a customer on a single visit to a store. An example association rule in this context is:

```
80% of the people who buy diapers and baby powder buy baby oil.
```

The number "80%" is referred to as the *confidence factor*, a measure of the predictive power of the rule. Here "diapers" and "baby powder" constitute the set of items on the left hand side (LHS) of the rule, and "baby oil" is the item on the right hand side (RHS) of the rule. Let us look at some applications of such discovered rules.

- Rules that have "Diet Coke" in the RHS

  Use: What should the store do to boost the sale of Diet Coke?
- Rules that have "Bagels" in the LHS

  Use: What products may be impacted if the store discontinues selling Bagels?
- Rules that have "Sausage" in the LHS and "Mustard" in the RHS

  Use: What items should be sold with sausage to promote the sales of mustard?

There can be any number of items on the LHS and RHS of association rules. The general model for association rules can be described as follows: Given a database of *transactions*, where each transaction consists of a number of *items*, such as retail transactions, a *pattern* is a particular set of items, such as {Milk Bread Eggs}. We can derive candidate rules from patterns. For example, given the pattern {Milk Bread Eggs}, a possible rule is:

```
Milk & Bread ==> Eggs
```

The *prevalence* of a pattern (and derived rules) is the number of transactions in which the pattern appears. The prevalence of the pattern {Milk Bread Eggs} is denoted by:

```
{Milk Bread Eggs}
```

We are interested only in patterns and rules that have some given *minimum prevalence* for two reasons.

1. A rule has business value only if a significant fraction of transactions support it.

   For example, it may be the case that everybody who buys “Caviar” buys “Vodka”. Thus, the rule

   ```
   Caviar ==> Vodka
   ```

   has 100% confidence. But if only a handful of people buy “Caviar”, the rule may be of limited value to the retailer.
2. A rule may not be statistically significant if a very small number of transactions support the rule.

   The rule may be observed due to chance, and it would not be prudent to make decisions based on such a rule.

The *confidence factor* described earlier is mathematically expressed as the ratio of the prevalence of the rule pattern to the prevalence of the LHS of the rule. For the rule

```
Milk & Bread ==> Eggs
```

the confidence factor is given by

$$\frac{|\texttt{Milk Bread Eggs}|}{|\texttt{Milk Bread}|}$$

The techniques needed to generate these association rules automatically require significant I/O bandwidth and computational power, and the data sets are large and need to be scanned fast. So for acceptable response time, the operations need to be parallelized, and the hardware and software must support these kinds of processing.

3.2 Profile Generation

Profile generation has many applications, including target marketing, attached mailings, credit approval and treatment-appropriateness determination.

Consider target marketing. Company X wishes to send out a promotional mailing. Company X can buy a general mailing list and would like to maximize the return on its mailing expenditure. Profile generation can help as follows. If the company has done a prior mailing, the responses to that mailing are available. The people on the previous mailing, their attributes, and their responses are input to the profile generator. The profile generator comes up with a characterization, or profile, of the people who responded to the previous mailing. This profile is then taken as a predictor of response to the current mailing. The mailing list is filtered to include only people who match the profile and the promotional mailing is sent to them.

As another example, consider credit or loan approval. Here, the company wishes to characterize the people who should be studied carefully before a loan or credit is granted to them. The input to the profile generation program is the prior lending or credit experience of the company. As with generating association rules, profile generation techniques can benefit from computational power and parallelism. In addition, large main memories can help to speed the process up significantly.

4.0 Characteristics of Data Mining

Figure 2, shows the data and information flow when the data mining techniques described in Section 3.0 are used. The parallel streams indicate that cost-effective, scalable parallelism is a critical technology for data mining.

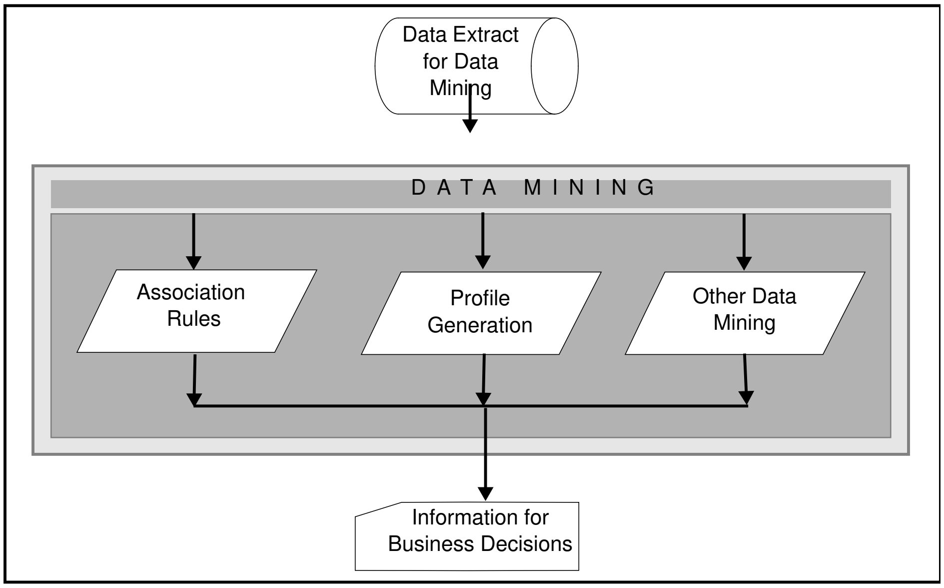

FIGURE 2 Data Mining

Data mining provides insights that are hard to come by using traditional techniques. The traditional techniques described in Section 1.0 complement the new data mining techniques. They cannot discover the kinds of information that were described before.

For example, the technique of mining for association rules allows the analyst to discover rules linking items in different categories. The techniques described in Section 1.0 work best at discovering rules within a certain category. The sheer number of possible rules between items in different categories is daunting, and only the new techniques have been successful in mastering this complexity.

Also, unlike techniques that mechanically aid an analyst in the discovery task, data mining techniques truly perform a kind of discovery. Typically, the results of data mining help to focus the analyst's attention. This, in turn, promotes the discovery of additional opportunities that are ripe for exploitation. However, discovery is hard, and data mining techniques may need to be adapted to the application domain.

Data mining analysis tends to be bottom-up, and the best techniques have been developed with an orientation towards large volumes of data. This is important in the context of the data warehouse, where a typical enterprise usually wishes to use as much of the collected data as possible to arrive at reliable conclusions and decisions.

5.0 Platforms for the Data Warehouse and Data Mining

Data mining technologies are characterized by intensive computations on large volumes of data. Significant processing power is critical, and parallelism is a key to enabling significant data mining. The systems should scale so that, as the demands for analysis grow, the system can be upgraded to provide the necessary analysis in a timely and cost-effective fashion.

Clearly, a balanced system architecture that supports I/O, computation and scaling in a cost-effective fashion is desirable. Systems used for updating operational data can be distinct from those used for the data warehouse, so the current enterprise investment in database and transaction platforms need not be wasted.

The requirements placed on a system for data mining are a superset of those for the data warehouse. Large main memories are necessary for good performance. Good memory management in the multiprocessor system is needed to ensure that thrashing at any level of the memory hierarchy does not result in inadequate performance.

The larger the volume of data that can be processed by data mining, the greater the confidence in the results. Hence, the highest capacity and performance systems are of interest in this area. Of course, price (as always) is a consideration.

There are two broad, alternative parallel architectures available now:

1. Massively parallel processing (MPP) systems

   Examples are Teradata from AT&T and SP-2 from IBM.
2. Symmetric multiprocessing (SMP) systems

   Examples are CHALLENGE servers from Silicon Graphics and servers from HP and Sequent.

Until now, doubts about their ability to meet the demand of data warehousing and mining have kept SMP systems from widespread use for these purposes. Today, however, the high performance, symmetric multiprocessing CHALLENGE line of servers from Silicon Graphics meets all these requirements. Unlike MPP systems, these servers are also excellent choices for general purpose use, thus helping to amortize the investment in system capacity.

The CHALLENGE servers provide the necessary system capabilities and can scale to meet increasing demands. The architecture and operating system are tuned to effectively support parallelism. Silicon Graphics has gained valuable experience with the demands placed by data-intensive applications and has tailored its systems to meet these demands.

The high-end CHALLENGE XL™ server can scale up to 36 processors (each 100 MIPS) with 16 gigabytes (GB) of main memory. It has 32 I/O channels where each I/O channel can support up to 20 MB per second. The server has sustained more than 500 MB per second of aggregate I/O bandwidth.[1] Using RAID technology, a single server can support more than 3.2 terabytes (TB) of data storage. Using 4 GB disk drives, a single server can have up to 2 TB of data storage, and using RAID technology, it can store more than 6 TB of data. Other CHALLENGE servers, such as the CHALLENGE L and CHALLENGE DM, are also suitable platforms for data mining.

The Oracle 7.1 RDBMS is a good choice as the data warehouse. The parallel loading capabilities of Oracle 7.1 make it possible to populate and refresh the repository in a timely fashion. The parallel query option enables quick retrieval of data from the warehouse for mining. The parallel query option can also be used for the complex queries that would be suggested by the data mining analysis. We have obtained a scaling factor of 11.5 with 16 CPUs running complex queries.[2]

Silicon Graphics has a close working relationship with Oracle Corporation to ensure that the system *as a whole* performs well. For example, a 200 GB database was installed at Silicon Graphics and the performance of loading, index creation and ad hoc querying using Oracle 7.1 was studied and tuned.[2]

Dramatic price/performance benefits can be gained by deploying the CHALLENGE servers for the data warehouse or data mining. As data points, Silicon Graphics CHALLENGE running Oracle 7 has demonstrated the best price/performance on the TPC-A and TPC-B benchmarks in the open systems market.[3]

Silicon Graphics has developed similar relationships with other database vendors, such as Sybase and Informix, and is committed to making the CHALLENGE servers the platforms of choice for data warehousing solutions based on those database systems.

6.0 Summary

Data mining is critical to the enterprise that wishes to exploit operational and other available data to improve the quality of decision making and gain critical competitive advantage. Companies that do not employ data mining may find themselves increasingly vulnerable in highly-competitive marketplaces.

The data mining techniques themselves require significant I/O bandwidth, large main memories, a lot of computational power, and scalable parallelism. Silicon Graphics CHALLENGE servers meet all these requirements; using a price/performance metric, CHALLENGE is the platform of choice for data mining the Open Systems market.

[1] See the University of Minnesota White Paper, *The Maximum Achievable Transfer Rate of a Silicon Graphics Challenge/Onyx Computer System*, by Thomas M. Ruwart.

[2] See the Silicon Graphics White Paper, *Using an SGI CHALLENGE Server to Build a 200 GB Data Warehouse*, by Bradley Eacker.

[3] See *TPC Benchmark™ A Full Disclosure Report for Silicon Graphics CHALLENGE XL Server and ORACLE7*, Silicon Graphics Document Number 007-2359-001.