Data Mining for Successful Healthcare Organizations

For successful healthcare organizations, it is important to empower the management and staff with data warehousing-based critical thinking and knowledge management tools for strategic decision-making. Many healthcare organizations struggle with the utilization of data collected through an organization's on-line transaction processing (OLTP) system, which is not integrated for decision-making and pattern analysis. Do you know what your top 10 paid procedures are? Do you know the DRGs that have the lowest reimbursement from Medicare/ Medicaid? Are you receiving timely and accurate reports for your analysis demands? Is your legacy system effective in identifying patterns to maximize your operational output? Are you drowning in data instead of sailing on the power of knowledge? This article focuses on data mining technology and its ability to uncover hidden and unexpected patterns in data for strategic decision-making in healthcare.

Healthcare Data

Healthcare data can be classified into four categories: *patient-centric data*, which is directly related to patients; *aggregate data*, which is based on performance and utilization/resource management data; *transformed-based data* for planning, clinical and management decision support; and *comparative data* for health services research and outcomes measurement. The data source is usually the on-line transaction processing (OLTP) system or operational support system (OSS) of an enterprise. Due to years of accumulation of disparate data, these systems are data rich but information poor. [6] With the adoption of data warehousing and data analysis/OLAP tools, an organization can make strides in leveraging data for better decision-making.

It should be noted here that these tools depend on the users to guide the data investigation process. Moreover, the tools require a predefined starting point, such as a hypothesis, a query, a procedure, or a program that dictates the nature of data analysis. These tools do not "uncover previously unknown business facts." [1] Therefore, there is a possibility that certain important information can remain untapped because no one knows of its existence.

Importance of Data Mining in Healthcare

Data mining, also known as database exploration and information discovery, brings the practice of information processing closer to providing the real answers organizations are seeking from their data. [1] In health care, there is massive data, and this data has no organizational value until converted into information and knowledge, which can help control costs, increase profits, and maintain high quality of patient care. Data mining provides automated pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods.

Are you drowning in data instead of sailing on the power of knowledge?

Without data mining, it is difficult to realize the full potential of data collected within an organization, as data under analysis is massive, highly dimensional, distributed, and uncertain. [1] For healthcare organizations to succeed, they must have the ability to capture, store and analyze data. [6] On-line analytical processing (OLAP) provides one way for data to be analyzed in a multidimensional capacity. [2] In order to understand the pivotal role data mining plays in a healthcare setting, it is important to first understand the role of a data warehouse, its various components, and how data mining methods can harness its power.

Introduction to Data Warehousing

A data warehouse can be considered a set of hardware and software components to analyze data for better decision-making. Bill Inmon, recognized as the father of the data warehouse concept, defines a data warehouse as a subject-oriented, integrated, time variant, non-volatile collection of data in support of management's decision making process. [3] Typically, a data

warehouse consists of a set of programs that extract data from the operational environment, a database that maintains data warehouse data, and systems that provide data to users. [3] A data warehouse enables you to tap into knowledge hidden in the massive amounts of data to understand business trends and make timely strategic decisions. If implemented correctly, a data warehouse can be the platform that contains an organization's data in a centralized and normalized form for deployment to users to enable them to perform simple reporting, complicated analysis, multidimensional analysis, population analysis, physician analysis and benchmarking, data mart deployment, clinical protocol development, reimbursement/cost analysis and much more. [6]

Data Warehousing & Analysis Tools

In order to leverage an organization's operational data to create strategic acumen, data warehousing can be supported by decision support tools, such as data marts, OLAP and data mining tools. A data mart is a complete "pie-wedge" of the overall data warehouse pie, a restriction of the data warehouse to a single business process or to a group of related business processes targeted toward a particular business group.[6] On-line analytical processing (OLAP) solutions provide a multi-dimensional view of the data found in relational databases, which store data in a two-dimensional format. [2]

Many healthcare organizations struggle with the utilization of data collected through an organization's on-line transaction processing (OLTP) system, which is not integrated for decision-making and pattern analysis.

OLAP makes it possible to analyze potentially large amounts of data with very fast response times, and provides the ability for users to "slice and dice" through the data, and drill down or roll up through various dimensions as defined by the data structure. [2] Data mining tools automate the analysis of data to find unexpected patterns or rules that management can use to tailor business operations. With data mining, the system researches the data and determines hidden patterns and associations, while the analyst determines what to do with the results. [1]

Data Mining – A Closer Look

In the early days of data warehousing, data mining was viewed as a "subset" of the activities associated with the warehouse. Today, while a warehouse may be a good source for the data to be mined, data mining is recognized as an independent activity that is paramount for success with data warehouse-based decision support systems. It should be noted that though warehousing and mining are related activities that reinforce each other, data mining does rely on a different set of data structures and processes, and caters to a different group of users than the typical warehouse. [5]

Data mining, also known as database exploration and information discovery, brings the practice of information processing closer to providing the real answers organizations are seeking from their data.

Data mining enables users to discover hidden patterns without a predetermined idea or hypothesis about what the pattern may be. The data mining process can be divided into two categories: discovering patterns and associations, and predicting future trends and behaviors using the patterns. [1] The power of data mining is evident as it can bring forward patterns that are not even considered by the user to search for. Hence, you have the answer to a question that was never asked. This is especially helpful when you are dealing with a large database where there may be an infinite number of patterns to identify. Interesting to note is the fact that "the more data in the warehouse, the more patterns there are, and the more data we analyze the fewer patterns we find." [5] What this means is that when there is richness of data and data patterns, it may be best to data mine different data segments separately, so that the influence of one pattern does not dilute the effect of another pattern in a large database. At this junction, the user should not fear the loss of certain indicators or patterns in a smaller segment. If a trend or pattern is applicable in an entire database, then it is also evident in a smaller segment. Thus, depending

on the circumstances and the nature of data in a warehouse, data mining can be generally performed on a segment of data that fits a business objective, rather than the whole warehouse.[5]

Types of Analysis Sessions in a Data Mine

In a data mine, analysis for pattern identification is done on a data segment, and the process of mining this data segment is called an *analysis session*. For example, management may want to predict the response to a flu vaccination initiative for the elderly of the community by analyzing previous such initiatives, or they may want to know the patient sample over various geographic regions, etc. If a user starts the analysis with a specific task, for instance, analyzing reimbursement by DRGs, then this analysis is a *structured analysis*. Structured analysis is generally done on a routine basis like analysis of quarterly costs, revenues, expenses, etc. to identify and forecast trends.

Data mining enables users to discover hidden patterns without a predetermined idea or hypothesis about what the pattern may be.

When the user wanders through a database without a specific goal, then he/she may discover interesting patterns that are not conceived previously. This type of analysis is termed *unstructured*.[5]

Data Mining Techniques

The following types of techniques are used in data mining: *decision trees*, *genetic algorithms*, *neural networks*, *predictive modeling*, *rule induction*, *fuzzy logic*, *k-NN*, and so forth. Considering the scope of this article, these techniques are discussed in a simple and general overview. Readers are encouraged to explore further depending on their interest. A *decision tree* is a tree-shaped structure that visually describes a set of rules that caused a decision to be made. For example, it can help determine the factors that affect kidney transplant survival rates. *Genetic algorithms* are optimization techniques that can be used to improve other data mining algorithms so that they derive the best model for a given set of data. For example, algorithms can help determine the optimal treatment plan for a particular diagnosis. *Neural networks* are non-linear predictive models that learn how to detect a pattern to match a particular profile through a training process that involves interactive learning, using a set of data that describes what you want to find. For example, they can help determine what disease a susceptible patient is likely to contract. Neural networks are good for clustering, sequencing, and predicting patterns, but their drawback is that they do not explain why they have reached a particular conclusion.

A data warehouse enables you to tap into knowledge hidden in the massive amounts of data to understand business trends and make timely strategic decisions.

Predictive modeling can be used to identify patterns, which can then be used to predict the odds of a particular outcome based upon the observed data. *Rule induction* is the process of extracting useful if/then rules from data based on statistical significance. *Fuzzy logic* handles imprecise concepts and is more flexible than other techniques. For example, it can help determine patients that are likely to respond to hospital/community initiatives to raise awareness to the cause of preventing AIDS. Finally, *k-NN* or *K-Nearest Neighbor* is a classic technique for discovering associations and sequences when the data attributes are numeric.[1]

Data Mining Architecture

In the early days, the "build a warehouse first, mine later" paradigm was generally followed by organizations. Classified by Parsaye as the “Sandwich Paradigm,” this "build first, think later" approach can be considered a "data dump paradigm" as in many cases it leads to the construction of a toxic data dump whose contents are not easily salvageable. A much better way is to "sandwich" the warehousing effort between two layers of mining -- thus understanding the data before warehousing it. This approach ensures that data mining complements a data warehouse.[4,5]

Data mining can exist in three basic forms: *above the warehouse* as a set of conceptual views; *beside the warehouse* as a separate repository; and *within the warehouse* as a distinct set of resources. Data mining *above the warehouse* provides a minimal architecture for analysis and identifying patterns, and is prescribed only when data mining is not a key objective for the warehouse. Data mining *beside the warehouse* allows data mining to be done in specific data segments with specific business objectives, and the exploratory nature of the data mining exercise does not interfere with the warehouse's routine processes of query and reporting. Data is moved from the warehouse to the mine, is restructured during the transformation, and then is analyzed.

When selecting a data mining tool or product, it is important to consider how well the product integrates with other components of a data warehouse.

Data mining *within the warehouse* works like a "republic within a republic" and is independent of the warehouse structures, but usually leads to loss of flexibility without any significant benefits. Thus, in most cases, the data mine *beside the warehouse* approach is prescribed. Stand-alone data mines can also exist in the case when it takes too long for an organization to build its corporate warehouse. In fact, the stand-alone data mine can be used later as a guide to design the warehouse architecture following the "Sandwich Paradigm." [4,5]

Requirements for a Data Mining Tool

When selecting a data mining tool or product, it is important to consider how well the product integrates with other components of a data warehouse. It is recommended that a two-way exchange interface with an OLAP tool is done, which will allow the OLAP tool to pass selected sub-sets of a data warehouse to the data mining tool to be mined for patterns that are difficult to observe using an OLAP tool, and the mining tool can forward its findings to the OLAP tool to verify the findings in a larger database. It is also important for a data mining tool to support rule conversion. Converting mined rules into SQL queries allows OLAP and other analysis tools to use the mined rules immediately for data segmentation purpose, and makes the mined rules available for reuse by other OLAP applications. [1]

The data mining process can be divided into two categories: discovering patterns and associations, and predicting future trends and behaviors using the patterns.

It is equally important for a data mining tool to have visualization capabilities. The success of data mining relies on how well the user can view, evaluate and act upon discovered patterns through the user interface tool. There must be a natural way for users to view the results of a data mining activity, and the user interface must integrate seamlessly into the user's current environment for ease-of-use. The power of automatic pattern discovery is enhanced if patterns can be viewed via charts and tables using color, position and size differentiators, etc.[1]

There are some other requirements that need to be considered to have success with a data-mining product. The data mining product architecture must account for scalability and run-time performance in its design, and use multiprocessors to take full advantage of performance-enhancing technology. The product must support a full range of data mining activities and techniques to enable users to see the same problem from many different angles for a thorough investigation. The tool must also support various possible data sources, both internal and external. Finally, it is important for the data mining tool to support data preparation activities, such as data cleansing, data description, data transformation, data sampling and data pruning. [1]

Conclusion

Data mining technology and techniques are invaluable for healthcare management and staff for analyzing clinical and financial data for strategic decision-

making. This technology can enable organizations to meet their business objectives for cost control, revenue generation, while maintaining high quality of care for the patients. Moreover, data mining tools and techniques can ease up the burden of management and staff faced with the growing pressures of lower payer reimbursements, increased labor costs, and a highly competitive healthcare marketplace. These tools can work as strategic weapons to convert data into information, creating business intelligence in the process.

The Shams Group (TSG)
A Knowledge Management Consulting & Software Company
www.shamsgroup.com

References

Ferguson, Mike. "Evaluating and Selecting Data Mining Tools." *InfoDB* 11:2, November 1997.
Hettler, Mark. "Data Mining Goes Multidimensional." *Healthcare Informatics*, March 1997.
Lambert, Bob. "Data Warehousing Fundamentals: What You Need to Know to Succeed." *Data Management Review*, March 1996.
Parsaye, K. "The Sandwich Paradigm for Data Warehousing and Mining." *Database Programming and Design*, April 1995.
Parsaye, K. "Data Mines for Data Warehouses." Information Discovery Inc., 1996.
Shams, K & Farishta, M. "Data Warehousing: Toward Knowledge Management." *Topics in Health Information Management* 21:3 (2001): 24-32.