International Journal of Engineering, Science and Metallurgy

(General article)

Data Mining: A Knowledge Discovery Approach

Rashmi[a]

[a]Om Institute of Technology & Management,Hisar

Accepted 17 July 2012, Available online 1Sept 2012

Abstract

We are in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and now we have sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. One of the wonderful & powerful means is data mining. In this paper, we present the detailed definition of data mining, terms used in data mining, A suitable architecture for data mining, how it works, its limitations & one of its interesting application & resulted outcome.

Keywords: *Data Mining, Knowledge Discovery, Data Cleaning, Data Warehouse, OLAP, TIA*

1. An introduction to data mining

Initially, the need and tendency to store the bulk of data has led to the creation of structured databases and database management systems (DBMS). The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also on tribute to recent massive gathering of all sorts of information. Today, we have farm ore information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data.[3] Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. [1]These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. Thus, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. In other words, Data mining, (the extraction of hidden predictive information from large databases,) is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. We can understand it as follows: Data mining can be performed on data represented in quantitative, textual or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the purchase of a sim number & recharging it), classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases), clustering(finding and visually documenting groups of previously unknown facts, such as

geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes). [2].

2. Various terms used in data mining:

(Data, Information, and Knowledge)
Data: Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions [4]
- Information: The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when. [4] Knowledge: Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts. [4] .Data Warehouses: The vast advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. A data warehouse as a storehouse, is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema.[2] Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Vast technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.[4].The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:
- Data cleaning: also known as data cleansing, it is a phase in which noise data an irrelevant data are removed from the collection.
- Data integration: at this stage, multiple data sources, often heterogeneous, maybe combined in a common source.
- Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results. [5] [6] [7]

These steps can be effectively understood by following figure:

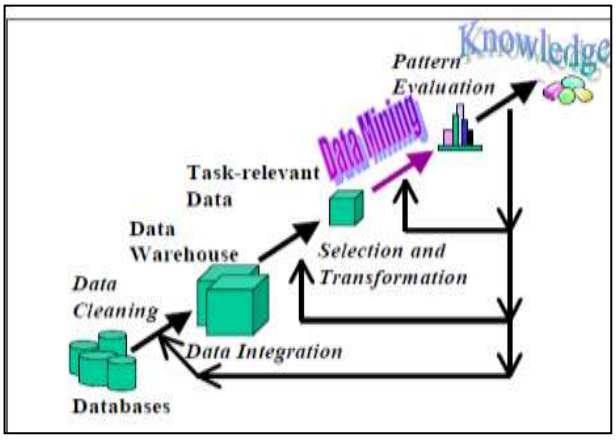

Fig.1 Data Mining is the core of Knowledge Discovery process [6].

It is possible to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

3. How data mining works:

To better understand how data mining works? First, we must know the architecture for data mining: The following figure shows architecture for advanced analysis in a large data warehouse:

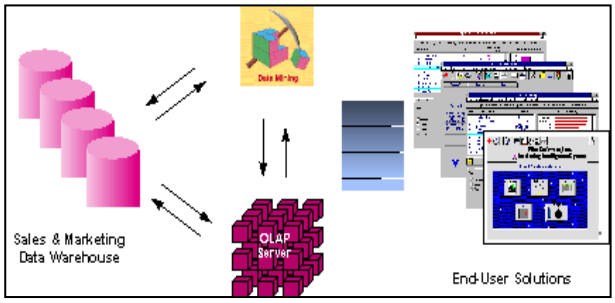

Fig.2 Integrated Data Mining Architecture [8]

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions. [9] Now, how data mining works: The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks:
Descriptive data mining tasks: that describes the general properties of the existing data Predictive data mining tasks: that attempt to do predictions based on inference on available data. The data mining functionalities and the variety of knowledge they discover are briefly as:

- Characterization: Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For example, one may want to characterize the Video Store customers who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the attribute oriented induction method can be used, for example, to carry out data summarization. Note that with a data cube containing summarization of data, simple OLAP operations fit the purpose of data characterization. [7]
- Discrimination: Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For example, one may want to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures.
- Classification: Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the Video Store managers could analyze the customers' behaviors vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels "safe", "risky" and "very risky". The classification analysis would generate a model that could be used to either accept or reject credit requests in the future.
- Prediction: Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values..
- Clustering: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many

clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

- Outlier analysis: Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.
- Evolution and deviation analysis: Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values. [10] [11] [12]

There is always a chance that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity an important attribute of a data mining system.

4. Limitations of data mining

While data mining products can be very powerful tools, they are not self-sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel related, rather than technology-related. [14]Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. Similarly, the validity of the patterns discovered is dependent on how they compare to "real world" circumstances. For example, to assess the validity of a data mining application designed to identify potential terrorist suspects in a large pool of individuals, the user may test the model using data that includes information about known terrorists. However, while possibly re-affirming a particular profile, it does not necessarily mean that the application will identify a suspect whose behavior significantly deviates from the original model. Another limitation of data mining is that while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. For example, an application may identify that a pattern of behavior,
such as the propensity to purchase airline tickets just shortly before the flight is scheduled to depart, is related to characteristics such as income, level of education, and Internet use. However, that does not necessarily indicate that the ticket purchasing behavior is caused by one or more of these variables. In fact, the individual's behavior could be affected by some additional variable(s) such as occupation (the need to make trips on short notice), family status (a sick relative needing care), or a hobby (taking advantage of last minute discounts to visit new destinations). [15]

5. Various uses of data mining

Data mining is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Using customer data collected over several years, companies can develop models that predict whether a customer is agood credit risk, or whether an accident claim may be fraudulent and should be investigated more closely. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine. Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases. Retailers can use information collected through affinity programs (e.g., shoppers' club cards, frequent flyer points, contests) to assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together. Companies such as telephone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor. [16]In the public sector, data mining applications were initially used as a means to detect fraud and waste, but they have grown also to be used for purposes such a measuring and improving program performance. It has been reported that data mining has helped the federal government recover millions of dollars in fraudulent medicare payments. [17] The Justice Department has been able to use data mining to assess crime patterns and adjust resource allotments accordingly. Similarly, the Department of Veterans Affairs has used data mining to help predict demographic changes in the constituency it serves so that it can better estimate its budgetary needs .Another example is the Federal Aviation Administration, which uses data mining to

review plane crash data to recognize common defects and recommend precaution measures. [18]

Recently, data mining has been increasingly cited as an important tool for homeland security efforts. Some observers suggest that data mining should be used as a means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. The initiatives that have attracted significant attention include the now-discontinued Terrorism Information Awareness (TIA) project [19] conducted by the Defense Advanced Research Projects Agency(DARPA).

6. What is TIA?

TIA purported to capture the "information signature" of people so that the government could track potential terrorists and criminals involved in "low-intensity/low-density" forms of warfare and crime. The goal was to track individuals through collecting as much information about them as possible and using computer algorithms and human analysis to detect potential activity. The project called for the development of "revolutionary technology for ultra-large all-source information repositories," which would contain information from multiple sources to create a "virtual, centralized, grand database." This database would be populated by transaction data contained in current databases such as financial records, medical records, communication records, and travel records as well as new sources of information. Also fed into the database would be intelligence data. A key component of the TIA project was to develop data-mining or knowledge discovery tools that would sort through the massive amounts of information to find patterns and associations. TIA aimed to fund the development of more such tools and data-mining technology to help analysts understand and even "preempt" future action.

7. Techniques used in TIA:

TIA is both a meta-search engine–querying many data sources at once–and a tool that performs pattern and link analysis that uncover hidden patterns and subtle relationships in data and infer rules that allow for the prediction of future results. Subject-based link analysis through utilization of the FBI's collection data sets, combined with public records on predicated subjects. Link analysis uses these data sets to find links between subjects, suspects, and addresses or other pieces of relevant information, and other persons, places, and things. This technique is currently being used on a limited basis by the FBI**.** The algorithm also uses pattern analysis as part of its queries that take a predictive model or pattern of behavior and search for that pattern in data sets starting from a predefined predictive model. TIA included several sub projects for mining different sets of data, such as: [22]

- Human Identification at a Distance
- Evidence Extraction and Link Discovery
- Scalable Social Network Analysis
- Translingual Information Detection, Extraction and Summarization
- Bio-Surveillance

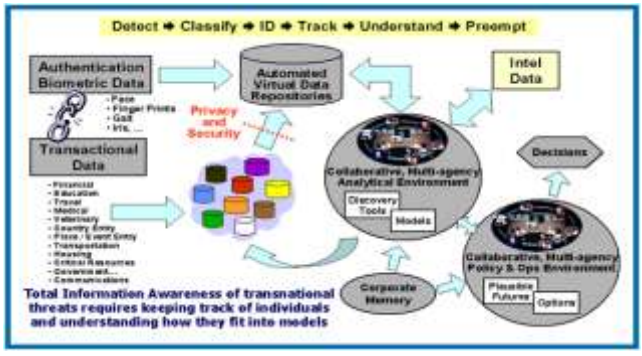

Fig.3 Total Information Awareness of transnational threats requires keeping track of individuals and understanding how they fit into module [21]

8. Results:

Although the investment in TIA was considerably huge and lots of efforts were put together in the data collection and mining process, the alleged ability of data-miners to discover hidden patterns and trends among disparate data-sets was native and lacked precision because so little is known about what patterns indicate terrorist activity; as a result, they are likely to generate huge numbers of false leads. Besides, Privacy concerns about mined or analyzed personal data also include concerns about the quality and accuracy of the mined data; the use of the data for other than the original purpose for which the data were collected without the consent of the individual; the protection of the data against unauthorized access, modification, or disclosure; and the right of individuals to know about the collection of personal information, how to access that information, and how to request a correction of inaccurate information. All these led to a failure in developing an efficient counter-terrorism prediction system and caused the complete termination of the TIA project. [23]

References

1. Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, Third Edition(Potomac, MD: Two Crows Corporation, 1999); Pieter Adriaans and Dolf Zantinge, Data Mining(New York: Addison Wesley, 1996).

2. [http://searchcrm.techtarget.com/gDefinition/0,294236,sid11_gci211901,00.html].
3. M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a data base perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.
4. J. Han and M. Kamber. Data Mining: Concepts and Techniques. MorganKaufmann, 2000.
5. W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, KnowledgeDiscovery in Databases: An Overview. In G. Piatetsky-Shapiro et al. (eds.),Knowledge Discovery in Databases. AAAI/MIT Press, 1991.
6. G. Piatetsky-Shapiro and W. J. FrawleyKnowledge Discovery in Databases.AAAI/MIT Press, 1991.
7. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy.Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
8. Osmar R. Zaïane, CMPUT690 Principles of Knowledge Discovery in Databases. 1999
9. Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 0471228524. OCLC 50055336.
10. T. Imielinski and H. Mannila. A database perspective on knowledgediscovery. Communications of ACM, 39:58-64, 1996.
11. G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining toknowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances inKnowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.
12. Adriaans, P., and Zantige, D. (1996) Data Mining. Harlow, UK: Addison-Wesley.
13. Abiteboul, S., Hull, R., and Vianu, V. (1995) Foundations of Databases. Reading, MA:Addison-Wesley.
14. Data Mining: Federal Efforts Cover a Wide Range of Uses, GAO Report GAO-04-548 (Washington: May 2004).
15. George Cahlink, "Data Mining Taps the Trends," Government Executive Magazine,October1,2000,[http://www.govexec.com/tech/articles/1000managetech.htm].
16. Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, Third Edition(Potomac, MD: Two Crows Corporation, 1999), p. 5; Patrick Dillon, Data Mining: Transforming Business Data Into Competitive Advantage and Intellectual Capital (Atlanta GA: The Information Management Forum, 1998), pp. 5-6.
17. George Cahlink, "Data Mining Taps the Trends," Government Executive Magazine,
18. October1,2000,[http://www.govexec.com/tech/articles/1000managetech.htm].
19. Data Mining: Federal Efforts Cover a Wide Range of Uses, GAO Report GAO-04-548 (Washington: May 2004).
20. This project was originally identified as the Total Information Awareness project until DARPA publicly renamed it the Terrorism Information Awareness project in May 2003.
21. Total Information Awareness (TIA), Electronic Privacy Information Center (EPIC)
22. http://epic.org/privacy/profiling/tia/
23. http://en.wikipedia.org/wiki/Information_Awareness_Office, http://dissidentvoice.org/2009/10/fbi-data- mining-programs-resurrect-total-information-awareness/