APPLICATION OF DATA MINING IN CHANGING TIMES AND ITS ROLE IN FUTURE

Faiz Malik,

Lecturer,
Department of Business Administration.
Jizan University. Jizan. KSA.

ABSTRACT

Data mining is growing and most important process of knowledge discovery in any possible concern of huge data base where knowledge is unearthed or gained by analyzing the data stored in very large repositories. An attempt is made in this paper to talk very briefly and systematically about the changes and historical development in data mining in changing times. It also gives importance about the current trend and future role of data mining.

***Keywords**: Artificial intelligence, Data mining, Machine Learning, Data mining trends, Data mining tools, patterns*

Introduction:

Data mining means searching for certain patterns within large sets of data, which creates a lot of possibilities for business managers and decision makers. By analyzing those patterns, better business decisions can be made in order to enable businesses to achieve greater financial and entrepreneurial success. Data mining has been garnering a significant amount of importance in recent years, creating a strong industrial impact. Based on this observation, it is evident that the future of data mining companies would be promising in the coming years. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time-consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

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 birth of a child and purchasing diapers), 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.

Data mining is finding increasing acceptance in science and business areas which need to analyze large amount of data to discover trends which they could not otherwise find (Data Mining Software http//www.dataminingsoftware. com). Data Mining can be seen as the confluence of multiple fields including statistics, machine learning, databases, pattern discovery and visualization etc. The various application areas of Data Mining are Life Sciences (LS), Customer Relationship Management (CRM), Web Applications, Manufacturing, Competitive Advantage, Intelligence, Retail, Finance, Banking, Computer, Network, Security, Monitoring, Surveillance, Teaching Support, Climate modeling, Astronomy, and Behavioral Ecology.

The objective of my present paper is to review trends of Data Mining and its relative areas from past to present. Keeping this objective in my mind I have organized this presentation in five sections:

1. Evolution of Data Mining
2. Current trends in Data Mining
3. Conclusion

Data Mining in Changing Times:

The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history. Data mining roots are traced back along three family lines: classical statistics, artificial intelligence, and machine learning (Data Mining Software http//www.datamining software.com).

Statistics:

Statistics are the foundation of most technologies on which data mining is built, e.g. regression analysis, standard distribution, standard deviation, standard variance, discriminant analysis, cluster analysis, and confidence intervals. All of these are used to study data and data relationships.

Artificial Intelligence:

Artificial Intelligence (AI) is one of the newest disciplines. It was formally initiated in 1956, when the name was coined, although at that point work had been under way for about five years. Along with modern genetics, it is regularly cited as the "field I would most like to be in" by scientists in other disciplines. A student in physics might reasonably feel that all the good ideas have already been taken by Galileo, Newton, Einstein, and the rest, and that it takes many years of study before one can contribute new ideas. AI, on the other hand, still has openings for a full-time Einstein. AI is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. Because this approach requires vast and robust computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. AI found a few applications at the very high-end scientific/government markets, but the required supercomputers of the era priced AI out of reach of virtually everyone else. Certain AI concepts which were adopted by some high-end commercial products, such as query optimization modules for Relational Database Management Systems (RDBMS)

Machine Learning:

Machine Learning (ML) is the union of statistics and AI. It could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals. Data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications. Data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning. These techniques are then used together to study data and find previously-hidden trends or patterns within.

Disciplines relevant to ML are:

- Artificial intelligence
- Bayesian methods
- Control theory
- Information theory
- Computational complexity theory
- Philosophy
- Psychology and neurobiology
- Statistics

When to use ML:

- When we do not know much about the problem
- When we have a lot of empirical data
- When we do not want to work hard

The commonly acceptable topics under ML are:

- Concept learning
- Decision trees
- Neutral networks
- Bayesian learning
- Ensemble methods
- Support Vector machines
- Instance-based methods
- Reinforcement learning
- Genetic algorithms
- Analytical learning
- Computational learning theory

Apart from the above three families discussed, data mining includes various other technologies or areas, e.g. pattern discovery, visualization, business intelligence etc.

Current Trends in Data Mining:

Data mining is formally defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The field of data mining has been growing rapidly due to its broad applicability, achievements and scientific progress, understanding. A number of data mining applications have been successfully implemented in various domains like fraud detection, retail health care, finance, telecommunication, and risk analysis.

The ever increasing complexities in various fields and improvements in technology have posed new challenges to data mining. The various challenges include different data formats, data from disparate locations, advances in computation and networking resources, research and scientific fields, ever growing business challenges etc. Advancements in data mining with various integrations and implications of methods and techniques have shaped the present data mining applications to handle the various challenges. The current trends of data mining applications are:

i) ***Terrorism Information Awareness*** (TIA): In the immediate aftermath of the September 11, 2001, terrorist attacks, many questions were raised about the USA's intelligence tools and capabilities, as well as the government's ability to detect other so-called "sleeper cells," if, indeed, they existed. One response to these concerns was the creation of the Information Awareness Office (IAO) at the Defense Advanced Research Projects Agency (DARPA) in January 2002. The role of IAO was "in part to bring together, under the leadership of one technical office director, several existing DARPA programs focused on applying information technology to combat terrorist threats." The mission statement for IAO suggested that the emphasis on these technology programs was to "counter asymmetric threats by achieving *total information awareness* useful for preemption, national security warning, and national security decision making." To that end, the TIA project was to focus on three specific areas of research, anticipated to be conducted over five years, to develop technologies that would assist in the detection of terrorist groups planning attacks against American interests, both inside and outside the country. Similar projects were also launched in European countries and rest of the world. This program, however, faced several problems:

a) The heterogeneity of database, the target database had to deal to deal with text, audio, image, and multimedia data.

b) Second problem was scalability of algorithms. The execution time increase as size of data (which is huge). For example, 230 cameras were placed in London to read number plates of vehicles. An estimated 40,000 vehicles pass camera every hour, in this way the camera must recognize 10 vehicles per seconds, which poses heavy loads on both hardware and software. (Transport for London 2004).

ii) **Bio-informatics and Health Care**: The second most important application trend, deals with mining and interpretation of biological sequences and structures. In healthcare, data mining is becoming increasingly popular. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. There is vast potential for data mining applications in healthcare. Generally, these can be grouped as the evaluation of treatment effectiveness; management of healthcare; customer relationship management; and detection of fraud and abuse. More specialized medical data mining, such as predictive medicine and analysis of DNA micro-arrays, lies outside the scope of this paper. Data mining tools are rapidly being used in finding genes regarding cure of diseases like Cancer and AIDS. Data mining applications in healthcare can have tremendous potential and usefulness. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry consider how data can be better captured, stored, prepared, and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications.

iii) **Web and Semantic Web:** Perhaps the most obvious instance of interaction between data mining and semantic web is web mining. According to a *Nature article* the World Wide Web doubles in size approximately every 8 months. There are approximately 20 million content areas in the Web. "85% of users use search engines to find information. Consumers use search engines to locate and buy goods or to research many decisions (such as choosing a vacation destination, medical treatment or election vote). However, the search engines are currently lacking in comprehensive and timeliness, and do not index sites equally. The current state of search engines can be compared to a phone book which is updated irregularly, is biased toward listing more popular information, and has most of the pages ripped out "*(*L. Giles and S. Lawrence). Though the Web is rich with information, gathering and making sense of this data is difficult because the document of the Web is largely unorganized. The biggest challenge in the next several decades is how to effectively and efficiently dig out a machine-understandable and queriable information and knowledge layer, called Semantic Web, from unorganized, human-readable Web data. Web is the hottest trend now, but it is unstructured. Data mining is helping web to be organized, which is called semantic web. The underlying technology is Resource Description Framework (RDF) which is used to describe resources. There are issues like combining all RDF statements and dealing with erroneous RDF statements. Data mining technologies are serving a lot to make the web, a semantic web.

iv) **Business Trend**: Today's business environment is more dynamic, so business must be able to react quicker, must be more profitable, and offer high quality services that ever before. Here, data mining serves as a fundamental technology in enabling customer's transactions more accurately, faster and meaningfully. Data mining techniques of classification, regression, and cluster analysis are used for in current business trends. Almost all of the current business data mining applications are based on the classification and prediction techniques for supporting business decisions, thus creating strong Business Intelligence (BI) system. Business data mining needs more enhancement in the design of data mining techniques to gain significant advantages in today's competitive global market place (E-Business). The Data mining techniques hold great promises for developing new sets of tools that can be

Table Showing the Data Mining Trends Comparative Statement

Data mining Trends	Algorithms/ Techniques employed	Data Formats	Computing Resources	Prime areas of applications
Past	Statistical, Machine Learning Techniques	Numerical data and structured data stored in traditional databases	Evolution of 4G PL and various related techniques	Business
Current	Statistical, Machine Learning, Artificial Intelligence, Pattern Reorganization Techniques	Heterogeneous data formats includes structured, semistructured and unstructured data	High speed networks, High end storage devices and Parallel, Distributed computing etc…	Business, Web, Medical diagnosis etc…
Future	Soft Computing techniques like Fuzzy logic, Neural Networks and Genetic Programming	Complex data objects includes high dimensional, high speed data streams, sequence, noise in the time series, graph, Multi-instance objects, Multirepresented objects and temporal data etc…	Multi-agent technologies and Cloud Computing	Business, Web, Medical diagnosis, Scientific and Research analysis fields (bio, remote sensing etc…), Social networking etc…

used to provide more privacy for a common man, increasing customer satisfaction, providing best, safe and useful products at reasonable and economical prices, in today's E-Business environment.

v) **Computing Resources**: The contemporary developments in high speed connectivity, parallel, distributed, grid and cloud computing has posed new challenges for data mining. The high speed internet connectivity has posed a great demand for novel and efficient data mining techniques to analyze the massive data which is captured of IP packets at high link speeds in order to detect the Denial of Service (DoS) and other types of attacks. Distributed data mining applications demand new alternatives in different fields, such as discovery of universal strategy to configure a distributed data mining, data placement at different locations, scheduling, resource management, and transactional systems etc. New data mining techniques and tools are needed to facilitate seamless integration of various resources in grid based environment. Moreover, grid based data mining has to focus seriously to address the data privacy, security and governance. Cloud computing is a great area to be focused by data mining, as the Cloud computing is penetrating more and more in all ranges of business and scientific computing. Data mining techniques and applications are very much needed in cloud computing paradigm.

vi) **Scientific Computing**: In recent years data mining has attracted the research in various scientific computing applications, due to its efficient analysis of data, discovering meaningful new correlations, patterns and trends with the help of various tools and techniques. More research has to be done in mining of scientific data in particular approaches for mining astronomical, biological, chemical, and fluid dynamical data analysis. The ubiquitous use of embedded systems in sensing and actuation environments plays major impending developments in scientific computing will require a new class of techniques capable of dynamic data analysis in faulty, distributed framework. The research in data mining requires more attention in ecological and environmental information analysis to utilize our natural environment and resources. Significant data mining research has to be done in molecular biology problems.

Comparative Statement:

The following table presents the comparative statement of various data mining trends from past to the future.

Conclusion:

In this paper a humble statement is made to understand the development that has taken place since the inception of data mining method.

The biggest challenge that data mining has been facing and it is going to face in the future is in the area of web and semantic web.

The linked data cloud now consists of 31 billion triples. This contains information about places, people, organisms, diseases, genes, medicines, and vast bibliographic data about books, music, television and movies. Can this data be mined? Certainly!

Major organizations are using RDF technology to support data mining today. Data mining is preceded a phase of data cleanup and integration; RDF, RDFS, SPARQL, SPIN and OWL are all suited for this task. Tools like Topbraid Composer and Virtuoso Open Link have powerful facilities to expose relational data in RDF form, and the W3C is developing standard ways to specify relational-RDF mapping.

Intelligent ETL systems will use RDF as an internal representation and will be able to insert selected RDF data into triple stores. They'll also be able to emit data in other formats such as XML or CSV file that can be fed into optimized analysis programs that use anything from FORTRAN to MapReduce.

Once we have data in graph format, there are many kinds of network analysis that are possible. For instance, PageRank, Hubs and Authorities, Centrality, Clustering, as well as very simple enumerations.

RDF data is suitable for this analysis. Useful graph mining algorithms can be implemented in SPARQL and OWL. Sometimes complex and powerful algorithms can be expressed in very simple code, although performance is not as good as custom-built software for the task. The bulk of data is difficult when you're dealing with a billion triples -- something you can get just by downloading a 6 gigabyte file.

However, we should never forget the fact that 60-80% of the labor effort in a successful data mining project goes into cleaning up data problems. If we want to do data mining on linked data, we will spend a lot of time cleaning up our source data -- semantic technology will cut this time down in the future, but for now we have to make the effort.

Data mining is a vast area and its future is very bright. It has several challenges as well. If careful steps are taken and the scientists and scholars of all over the world are working carefully, it can solve any problem and give the very accurate trend and data base in all possible areas. Though very few areas are discussed here in this presentation, but they are very pertinent and useful for the multiple stakeholders.

Abbreviations:

Active Data Objects (ADO)
Artificial Intelligence (AI)
Business Intelligence (BI)
CRoss-Industry Standard Process for Data Mining (CRISP-DM)
Customer Relationship Management (CRM)
Data Mining (DM)
Data Mining and Knowledge Discovery (DMKD)
Data Mining Group (DMG)
DMM (Data Mining Model)
Defense Advanced Research Projects Agency (DARPA)
Denial of Service (DoS)
Extensible Markup Language (XML)
Information Awareness Office (IAO)
Knowledge Discovery (KD)
Knowledge Discovery in Databases (KDD)
Life Sciences (LS)
Machine Learning (ML)
On-Line Analytical Processing (OLAP)
OLE DM for Data Mining (OLE DB-DM)
Predictive Model Markup Language (PMML)
Relational Database Management Systems (RDBMS)
Resource Description Framework (RDF)
Secure Multiparty Computation (SMC)
Terrorism Information Awareness (TIA)
Universal Description Discovery and Integration (UDDI)

References:

[1] For a more technically-oriented definition of data mining, See http://searchcrm.techtarget.com /gDefinition/0, 294236,sid11_gci211901,00.html

[2] Han, J., & Kamber, M. 2001. Data Mining concepts and Techniques", Morgan-Kaufman Series of Data Management Systems: San Diego: Academic Press.

[3] Stuart J. Russell and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*, Prentice Hall, New Jersey.

[4] Kumar, Dharminder & Deepak Bhardwaj. 2011. "Rise of Data mining: Current and Future Application Areas", *IJCSI International Journal of Computer Science* Issues, pp. 256- 260, Vol. 8, Issue 5, No.1, September 2011; http://www.unc.edu/~ xluan/258/datamining.html#history

[5] Fayadd, U., Piatesky-Shapiro, G., and Smyth, P. 1996. *From Data Mining To Knowledge Discovery in Databases*, AAAI Press, Massachusetts.

[6] M. Venkatadari, Lokanatha C. Reddy. 2011. "A Review of Data Mining from Past to Future", *International Journal of Computer Applications*, pp. 19-22, Vol. 15, No. 7.

[7] Seifert Jeffrey W. 2004. "Data Mining: An Overview", *CRS Report for Congress*; Huysmans, Baesens, Martens, Denys and Vanthienen. 2005. "New Trends in Data Mining", *Tijdschrift voor Economie en management*, Vol. L. No. 4.

[8] Mark, J., Embrechts. 2005. "Introduction to Scientific Data Mining: Direct Kernel Methods & Applications", *Computationally Intelligent Hybrid Systems: The Fusion of Soft Computing and Hard Computing*, Wiley, New York, pp. 317-365

[9] Han, J., & Kamber, M. 2001. *Data mining: Concepts and techniques, Morgan-Kaufman Series of Data Management Systems*, San Diego: Academic Press

[10] M., Venkatadri & Lokanatha C. Reddy. 2011. "A Review of Data Mining from past to Future", *International Journal of Computer Applications*, Vo. 15, No. 7., pp.19-22.

******