IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 15-20

Data Mining: Past, Present and Future

Dr. Dhiraj N. Shembekar[*1],Prof. Sudhir Juare[2]
[*1]*Department. of Computer Science, G.H. Raisoni Institute ofInformation Technology, Nagpur*
[2]*Department. of Computer Science, G.H. Raisoni Institute of Information Technology, Nagpur*

Abstract: *Knowledge has played a significant role in every sphere of human life. To acquire knowledge we have to analyze the unlimited data that is available to us in various formats in the form of databases. Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. Data mining roots are traced back along three family lines: classical statistics, artificial intelligence, and machine learning.*
The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history. Term "Knowledge Discovery in Databases" for Information Harvesting, Information Discovery, Knowledge Extraction, etc introduce by Gregory Piatetsky-Shapiro (1989) and this term became more popular in AI and Machine Learning Community. Presently Data Mining working on tremendous application in several areas like Business, Medical science and Sciences, engineering, Psychology and much more. But still there are several challenges for data mining for better services like scaling, algorithms, security etc. which will be the future opportunities for researchers.
Keywords: *Knowledge, Data Mining*

I. Introduction

Rapid advances in data collection and storage technology have enabled organizations to accumulate vast amounts of data. However, extracting useful information has proven extremely challenging. Often, traditional data analysis tools and techniques cannot be used because of the massive size of a data set. Sometimes, the non-traditional nature of the data means that traditional approaches cannot be applied even if the data set is relatively small. In other situations, the questions that need to be answered cannot be addressed using existing data analysis techniques, and thus, new methods need to be developed.

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. It has also opened up exciting opportunities for exploring and analyzing new types of data and for analyzing old types of data in new ways.**Data mining**is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.

Knowledge has played a significant role in every sphere of human life. To acquire knowledge we have to analyze the unlimited data that is available to us in various formats in the form of databases. We can analyze this data and find hidden information with the support of data mining. Data mining refers to the process or method that extracts interesting knowledge from large amounts of data. Data mining have number of applications and these applications have enhanced the various fields of human life including business, education, social media medical, scientific etc.

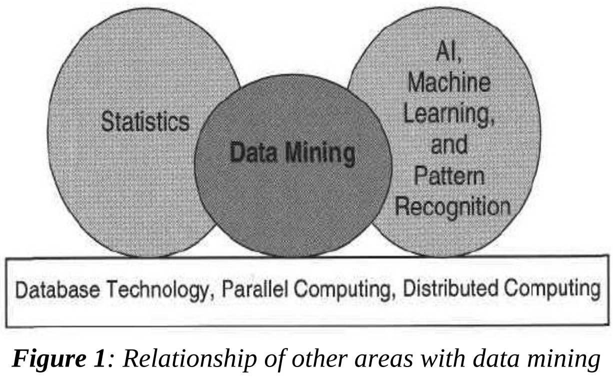

Figure 1*: Relationship of other areas with data mining*

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.

II. Derivation of Data Mining:

In the 1960s, statisticians used terms like “Data Fishing” or “Data Dredging” to refer to what they considered the bad practice of analyzing data without an a-priori hypothesis. The term “Data Mining” appeared around 1990 in the database community. At the beginning of the century, there was a phrase “database mining”™, trademarked by HNC, a San Diego-based company (now merged into FICO), to pitch their Data Mining Workstation; researchers consequently turned to “data mining”. Other terms used include Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, etc. Gregory Piatetsky-Shapiro coined the term “Knowledge Discovery in Databases” for the first workshop on the same topic (1989) and this term became more popular in AI and Machine Learning Community. However, the term data mining became more popular in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably.

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. Statistics are the foundation of most technologies on which data mining is built, e.g. regression analysis, standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis, and confidence intervals. All of these are used to study data and data relationships. Artificial intelligence, or AI, which is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. 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 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.

The term "data mining" is in fact a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

Researchers from different disciplines began to focus on developing moreefficient and scalable tools that could handle diverse types of data. This work,which culminated in the field of data mining, built upon the methodology andalgorithms that researchers had previously used. In particular, data miningdraws upon ideas, such as sampling, estimation, and hypothesis testingfrom statistics and search algorithms, modeling techniques, and learningtheories from artificial intelligence, pattern recognition, and machine learning.Data mining has also been quick to adopt ideas from other areas, includingoptimization, evolutionary computing, information theory, signal processing,visualization, and information retrieval.A number of other areas also play key supporting roles. In particular,database systems are needed to provide support for efficient storage, indexing,and query processing. Techniques from high performance (parallel) computingare often important in addressing the massive size of some data sets.Distributed techniques can also help address the issue of size and are essentialwhen the data cannot be gathered in one location.

III. Current Data Mining In Different Area With Problems Review

Scalability, today’s first problem presents in data mining.Computer Science researcher Alan Demers and Gehrke are workingwith Jim Cordes of the AstronomyDepartment on the design andimplementation of an analysisinfrastructure for a new censusof pulsars in the Milky WayGalaxy. The data will be collectedat the Arecibo Observatory inPuerto Rico. “The data rates andprocessing requirements for thepulsar survey are truly astronomical,”says Gehrke. The total rawdata, which will take three to fiveyears to acquire, will be aboutone petabyte —14 terabytes ofdata will arrive every two weeksvia “Fed-Ex-Net” on USB diskpacks, requiring the processing of one TB of dataper day. A recent $2M research infrastructure awardhas allowed the team to build the necessary computinginfrastructure at the Cornell Theory Center.

A second problem is to mine data with missingor wrong entries. Computer Science professor Rich Caruana andresearcher MirekRiedewald are working with scientistsfrom the Cornell Lab of Ornithology on analyzinglarge citizen-produced datasets. Every year, tensof thousands of volunteers report sightings of birdsto the Cornell Lab of Ornithology, creating one ofthe largest and longest-running resources of environmentaltime-series data in existence. Its analysiscould reveal long-term changes in ecosystems due tohuman intervention; for example changes in farmingpractices have been shown to affect bird abundanceover time. But mining the data is challenging. Volunteersoften leave some entries in bird report formsempty, novice observers may confuse bird species,and other variables such as habitat, weather, humanpopulation, climate, and geography have to beconsidered when estimating the true abundance of aspecies. “Compensating for bias in the collected datais a major challenge, and each observation could bedifferently biased,” says Caruana.

Steps of Data Mining to Obtain Knowledge

A third problem is the enormous complexity oftoday’s databases. For example, consider the Web.CS professors Bill Arms, Gehrke, Dan Huttenlocher,Jon Kleinberg, and Jai Shanmugasundaramare building a testbed that will enable the study oftemporal dynamics of the Web over time. The team will obtain the 40 billion Web pages archived bythe Wayback Machine, the time machine of the Internet.The team will also receive new 20–terabytesnapshots of Web crawls every two months. Thiscollection will enable the research community, forthe fi rst time, to evaluate models of Web growth andevolution at a wide range of different time scales.“The combination of content, link structure, andtemporal evolution creates an immensely complexdataset,” says Arms. “With this data and associateddata-mining tools, we will be able to tackle reallybig questions, for example how new technologies,opinions, fads, fashions, norms, and urban legendsspread over time.”“The beauty of working in this area is that you havediscovery at two levels,” says Gehrke. “You developinteresting new computer science methods, and youfind nuggets by applying these to real datasets.”

IV. Present Scenario

Neural Networks. Neural networks are systems inspired by the human brain. A basic example isprovided by a back propagation network which consists of input nodes, output nodes, andintermediatenodes called hidden nodes. Initially, the nodes are connected with random weights. During the training,a gradient descent algorithm is used to adjust the weights so that the output nodes correctly classify datapresented to the input nodes. The algorithm was invented independently by several groups ofresearchers.

Tree-based Classifiers. A tree is a convenient way to break a large data sets into smaller ones. Bypresenting a learning set to the root and asking questions at each interior node, the data at the leaves canoften be analyzed very simply. For example, a classifier to predict the likelihood that a credit cardtransaction is fraudulent may use an interior node to divide a training data set into two sets, dependingupon whether or not five or fewer transactions were processed during the previous hour. After a series ofsuch questions, each leaf can be labeled fraud/no-fraud by using a simple majority vote. Tree basedclassifiers were independently invented in information theory, statistics, pattern recognition and machinelearning.

Graphical Models and Hierarchical Probabilistic Representations. A directed graph is a good means oforganizing information about qualitative knowledge about conditional independence and causalitygleamed from domain experts. Graphical models generalize generalize Markov models and hiddenMarkov models, which have proved themselves to be a powerful modeling tool. Graphical models wereindependently invented by computational probabilists and artificial intelligence researchers studyinguncertainty.

Ensemble Learning. Rather than use data mining to build a single predictive model, it is often better tobuild a collection or ensemble of models and to combine them, say with a simple, efficient votingstrategy. This simple idea has now been applied in a wide variety of contexts and applications. In somecircumstances, this technique is known to reduce variance of the predictions and therefore to decreasethe overall error of the model.

Linear Algebra. Scaling data mining algorithms often depends critically upon scaling underlyingcomputations in linear algebra. Recent work in parallel algorithms for solving linear system

andalgorithms for solving sparse linear systems in high dimensions are important for a variety of datamining applications, ranging from text mining to detecting network intrusions.
Large Scale Optimization. Some data mining algorithms can be expressed as large-scale, oftennon-convex, optimization problems. Recent work has provided parallel and distributed methods forlarge-scale continuous and discrete optimization problems, including heuristic search methods forproblems too large to be solved exactly.

High Performance Computing and Communication. Data mining requires statistically intensiveoperations on large data sets. These types of computations would not be practical without the emergenceof powerful SMP workstations and high performance clusters of workstations supporting protocols forhigh performance computing such as MPI and MPIO. Distributed data mining can require moving largeamounts of data between geographically separated sites, something which is now possible with theemergence of wide area high performance networks.

Databases, Data Warehouses, and Digital Libraries. The most time consuming part of the data miningprocess is preparing data for data mining. This step can be stream-lined in part if the data is already in adatabase, data warehouse, or digital library, although mining data across different databases, forexample, is still a challenge. Some algorithms, such as association algorithms, are closely connected todatabases, while some of the primitive operations being built into tomorrow's data warehouses shouldprove useful for some data mining applications.

Visualization of Massive Data Sets. Massive data sets, often generated by complex simulation programs,required graphical visualization methods for best comprehension. Recent advances in multi-scalevisualization allow the rendering to be done far more quickly and in parallel, making these visualizationtasks practical.

V. Present Area Of Applications

The discipline of data mining is driven in part by new applications which require new capabilities notcurrently being supplied by today's technology. These new applications can be naturally divided intothree broad categories.

a. **Business & E-commerce Data.** Back-office, front-office, and network applications produce largeamounts of data about business processes. Using this data for effective decision making remains afundamental challenge.
b. **Scientific, Engineering & Health Care Data.** Scientific data and meta-data tend to be morecomplex in structure than business data. In addition, scientists and engineers are making increasinguse of simulation and of systems with application domain knowledge.
c. **Web Data.** The data on the web is growing not only in volume but also in complexity. Web datanow includes not only text and image, but also streaming data and numerical data.In this section, we describe several such applications from each category.

Business Transactions: Today, businesses are consolidating and more and more businesses have millionsof customers and billions of their transactions. They need to understand risks (Is this transactionfraudulent? Will this customer pay their bills?) and opportunities (What is the expected profit of thiscustomer? What product is this customer most likely to buy next?).

Electronic Commerce: Not only does electronic commerce produce large data sets in which the analysisof marketing patterns and risk patterns is critical, but unlike some of the applications above, it is alsoimportant to do this in real or near-real time, in order to meet the demands of on-line transactions.
Genomic Data: Genomic sequencing and mapping efforts have produced a number of databases whichare accessible over the web. In addition, there are also a wide variety of other on-line databases,including those containing information about diseases, cellular function, and drugs. Findingrelationships between these data sources, which are largely unexplored, is another fundamental datamining challenge. Recently, scalable techniques have been developed for comparing whole genomes.

Sensor Data: Satellites, buoys, balloons, and a variety of other sensors produce voluminous amounts ofdata about the earth's atmosphere, oceans, and lands. A fundamental challenge is to understand therelationships, including causal relationships amongst this data. For example, do industrial pollutantsaffect global warming? There are also large terabyte to petabyte data sets being produced by sensors andinstruments in other disciplines, such as astronomy, high energy physics, and nuclear physics.

Simulation Data: Simulation is now accepted as a third mode of science, supplementing theory andexperiment. Today, not only do experiments produce huge data sets, but so do simulations. Data mining,and more generally data intensive computing, is proving to be a critical link between theory, simulation,and experiment.

Health Care Data: Health care has been the most rapidly growing segment of the nation's GDP for sometime. Hospitals, health care organizations, insurance companies, and the federal government have largecollections of data about patients, their health care problems, the clinical procedures used, their costs,and the outcomes. Understanding relationships in this data is critical for a wide variety of problems,ranging from determining what procedures and clinical protocols are most effective to how best todeliver health care to the most people in an era of diminishing resources.
Multi-media Documents: Few people are satisfied with today's technology for retrieving documents onthe web, yet the number of documents and the number of people accessing these documents is growingexplosively. In addition, it is becoming easier and easier to archive multi-media data, includingaudio, images, and video data, but harder and harder to extract meaningful information from the archivesas the volume grows.

The Data Web: Today the web is primarily oriented toward documents and their multi-media extensions.HTML has proved itself to be a simple, yet powerful language for supporting this. Tomorrow thepotential exists for the web to prove equally important for working with data. The Extensible MarkupLanguage (XML) is an emerging language for working with data in networked environments. As thisinfrastructure grows, data mining is expected to be a critical enabling technology for the emerging dataweb.

VI. Future Challenges

A.Scaling data mining algorithms. Most data mining algorithms today assume that the data fits intomemory. Although success on large data sets is often claimed, usually this is the result of sampling largedata sets until they fit into memory. A fundamental challenge is to scale data mining algorithms as
1. The number of records or observations increases;
2. The number of attributes per observation increases;
3. The number of predictive models or rule sets used to analyze a collection of observations increases;
4. And, as the demand for interactivity and real-time response increases.

Not only must distributed, parallel, and out-of-memory versions of current data mining algorithms bedeveloped, but genuinely new algorithms are required. For example, association algorithms today cananalyze out-of-memory data with one or two passes, while requiring only some auxiliary data be kept inmemory.

B.Extending data mining algorithms to new data types. Today, most data mining algorithms work withvector-valued data. It is an important challenge to extend data mining algorithms to work withother data types, including 1) time series and process data, 2) unstructured data, such as text, 3)semi-structured data, such as HTML and XML documents, 4) multi-media and collaborative data, 5)hierarchical and multi-scale data, and 6) and collection-valued data.

C.Developingdistributed data mining algorithms. Today most data mining algorithms require bringingall together data to be mined in a single, centralized data warehouse. A fundamental challenge is todevelop distributed versions of data mining algorithms so that data mining can be done while leavingsome of the data in place. In addition, appropriate protocols, languages, and network services arerequired for mining distributed data to handle the meta-data and mappings required for miningdistributed data. As wireless and pervasive computing environments become more common, algorithmsand systems for mining the data produced by these types of systems must also be developed.

D.Ease of Use. Data mining today is at best a semi-automated process and perhaps destined to alwaysremain so. On the other hand, a fundamental challenge is to develop data mining systems which areeasier to use, even by casual users. Relevant techniques include improving user interface, supportingcasual browsing and visualization of massive and distributed data sets, developing techniques andsystems to manage the meta-data required for data mining, and developing appropriate languages andprotocols for providing casual access to data. In addition, the development of data mining andknowledge discovery environments which address the process of collecting, processing, mining, andvisualizing data, as well as the collaborative and reporting aspects necessary when working with dataand information derived from it, is another important fundamental challenge.

E.Privacy and Security. Data mining can be a powerful means of extracting useful information fromdata. As more and more digital data becomes available, the potential for misuse of data mining grows.

Afundamental challenge is to develop privacy and security models and protocols appropriate for datamining and to ensure that next generation data mining systems are designed from the ground up toemploy these models and protocols.

VII. Conclusion

Data mining is most valuable technique to obtain the hidden knowledge from database. The Data mining initially started for finding the hidden information, but later on it moves toward the finding pattern. Hidden information just gives the unknown info about the entity or object, but by pattern understanding through the data mining possible to forecast the future. Current data mining process and technique are very modern combination of statistical tools with AI. But still some problems in data mining give the opportunities to improve it. Data mining is blessing for the business, science and technology.

References

[1]. https://www.information-age.com/importance-data-mining-123469819/
[2]. https://www.cs.cornell.edu/gries/40brochure/pg18_19.pdf
[3]. http://iranarze.ir/wp-content/uploads/2018/10/9325-English-IranArze.pdf
[4]. https://pdfs.semanticscholar.org/8a60/b0 82aa758c317e9677beed7e7776acde5e4c.pdf
[5]. http://sites.bu.edu/phenogeno/files/2014/06/grossman98-Data-minin-research-opportunities.pdf
[6]. https://www.researchgate.net/p ublication/304808426_Data_Minig_in_Education/download
[7]. https://www.researchgate.net/publication/264458763_Application_of_Data_Mining_in_Educational_Database_for_Predicting_Behavioural_Patterns_of_the_Students/download
[8]. https://airccse.com/oraj/papers/1114oraj04.pdf
[9]. https://en.wikipedia.org/wiki/Data_mining
[10]. http://rastives.weebly.com/uploads/5/6/6/8/5668001/history_of_data_mining.pdf
[11]. Introduction to Data Mining by Pang.Ni Ng Tan, Michael Steinbach, Vi Pi N Ku Mar, ISBN 0-321-42052
[12]. Data Warehousing by CSR Prabhu, PHI Publication