iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** XI **Month of publication:** November 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Data Mining for Real Time Applications and Security Issues

[1]S Lavanya Reddy, [2] A V S M Adiseshu, [3]P Hasitha Reddy, [4]Roshini K

[1]*Associate Professor,* [2]*Assistant Professor,* [3]*Assistant Professor,* [4]*Assistant Professor*
[1,2,3,4] *Department of Computer Science & Engineering, Sree Dattha Institute of Engineering & Science*

Abstract - Data mining is used to extract knowledge from huge amount of the data Today, Data mining helps different organizations focus on customer's behavior patterns. The research scope of data mining extended in various fields. This paper, discusses the concept of data mining, important issues and applications. So there comes the need of powerful and most importantly automatic tools for uncovering valuable slots of organized information from tremendous amount of data. Considering social networking site or a search engine, they receive millions of queries every day. Firstly, the Database Management Systems evolved to handle the queries of similar types. Then the approach was modified to advanced Database management system, Data warehousing and Data mining for advance data analysis and web based databases. Data mining has immensely penetrated in each and every field of day to day life
Keywords — Data Mining, Data Analysis, Data Base, KDD, Information, Application Areas, Mining Techniques

1. INTRODUCTION

A proverb says that, "We are living in the era of information". But the reality is somewhat different, as raw data cannot be used, it needs to be refined and converted to information that is compatible with the raw data. Figure 1 describes the steps in Knowledge discovery in data bases. The steps are Data mining is the process of discovering Meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques". Data mining sometimes called data or knowledge discovery. 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.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information the patterns, associations, or relationships among all this *Data* can provide *information.* 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. As an Advancement of information technology in various fields of human life has increased to the large amount of data storage in various ways like records, documents, images, sound recordings, videos, scientific data, and many new data formats For better decision making, the large repositories data collected from different resources require proper mechanism of extracting knowledge from the databases. Knowledge discovery in databases (KDD), often called data mining, extracting information and patterns from data in large data base. The core functionalities of data mining are applying various techniques to identify nuggets of information of decision making knowledge in bodies of data. From the last decades, data mining and knowledge discovery applications have important significance in decision making and it has become an essential component in various organizations and fields. The field of data mining has been increased day by day in the areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc.

II. INTRODUCTION

According to the Gartner Group, "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques." "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner".

"Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics,

databases, and visualization to address the issue of information extraction from large data bases".

III. WHY USE DATA MINING?

Data mining is to extract information from large amount of a data base. Aristotle Onassis said "The secret of success is to know something that nobody else knows".
There are two main reasons to use data mining as a rapidly increase demands of data. These are:

A. Too much data and too little information.
B. There is a need to extract useful information from the data and to interpret the data.
C. Competitive Pressure in business
D. Quality, Timely Response etc.

IV. HISTORY OF DATA MINING

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, Artificial intelligence, Relational Database Management Systems (RDBMS), Machine learning. 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.

V. ISSUES OF DATA MINING

One of the key issues raised by data mining technologies is not a business or technological one, but social one. Some of the issues are address below:

A. Security and social issues

Today, Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. When data is collected for customer profiling, user behavior understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control.

B. User interface issues

The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge.
The major issues related to user interfaces and visualization is "screen real-estate", information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels.

C. Mining methodology issues

These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs (when known), the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover, different approaches may suit and solve user's needs differently

D. Performance issues

Many artificial intelligence and statistical methods exist for data analysis and interpretation. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are incremental updating, and parallel programming. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data mining results when new data becomes available without having to re-analyze the complete dataset.

E. Data source issues

There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate. If the spread of database management systems has helped increase the gathering of information, the advent of data mining is certainly encouraging more data harvesting. The current practice is to collect as much data as possible now and process it, or try to process it, later.

VI. TECHNIQUES USED IN DATA MINING

Several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns. We will briefly examine them with example to have a good overview of them.

A. Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in reservation systems analysis to identify in which area customers frequently make reservations. Based on this data businesses can set up corresponding reservation counters in that area to sell more tickets and make more profit.

B. Classification

Classification is based on machine learning. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. Basically classification is used to categorize each item in a set of data into one of predefined set of classes or groups]. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay".

C. Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. Consider library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic

without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

D. Prediction

It is one of a data mining techniques that discover relationship between independent variables and relationship between dependent and independent variables [3]. For instance, prediction technique can be used in Library to predict books that need to be purchased for the future if we assume that the courses offered by a university are constant. Courses are independent variable, and books could be a dependent variable.

E. Sequential Patterns

Sequential patterns analysis is one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

F. Discrimination

Data discrimination produces what are called discriminate 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

VII. APPLICATIONS OF DATA MINING

Data mining is a process that analyzes the large amount of data to find the new and hidden information that improves business efficiency. Various industries have been adopting data mining to their mission-critical business processes to gain competitive advantages. The data mining applications in sale/marketing, finance, health care and insurance, transportation and medicine and many other sectors of day to day life are remarkable. But some other distinct applications of data mining are listed below:

A. Computer Security

It concentrates heavily on the use of data mining in the area of intrusion detection. The reason for this is twofold. First, the volume of data dealing with both network and host activity is so large that it makes it an ideal candidate for using data mining techniques. Second, intrusion detection is an extremely critical activity. This book also addresses the application of data mining to computer forensics. This is a crucial area that seeks to address the needs of law enforcement in analyzing the digital evidence

B. Bioinformatics

Developments in genomics and proteomics have generated a large amount of biological data in the near past. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data. Analyzing large biological data sets requires making sense of the data by Inferring structure or generalizations from the data. Specific applications in this section of data mining are protein structure prediction, gene classification, cancer classification etc. Hence we can say that there is potential increase in the interaction between data mining and bioinformatics.

C. Telecommunications Industry

The telecommunications industry was one of the first to adopt data mining technology. This is most likely because telecommunication companies routinely generate and store enormous amounts of high-quality data, have a very large customer base, and operate in a rapidly changing and highly competitive environment. Telecommunication companies utilize data mining to improve their marketing efforts, identify fraud, and better manage their telecommunication networks [6]. However, these

companies also face a number of data mining challenges due to the enormous size of their data sets, the sequential and temporal aspects of their data, and the need to predict very rare events—such as customer fraud and network failures—in real-time.

D. Customer Relationship Management

CRM can be defined as the process of predicting customer behavior and selecting actions to influence that behavior for the benefit of the company [8]. What marketers want is nothing but "Increasing customer revenue and customer profitability and keeping the customers for a longer period of time." The solution is to apply data mining. Data mining techniques can be of immense help to the organization in solving business problems by: Finding patterns, associations and correlations which are hidden in the business information stored in the databases.

E. Banking

Apart from execution of business processes, the creation of knowledge base and its utilization for the benefit of the organization is becoming a strategy tool to compete. The banking sector has started realizing the need of the techniques like data mining which can help them to compete in the market. Since 1990's the whole concept of banking has been shifted to centralized databases, online transactions and ATM's all over the world, which has made banking system technically strong and more customer oriented. In the present day environment, the huge amount of electronic data is being maintained by banks around the globe. The huge size of these data bases makes it impossible for the organizations to analyze these data bases and to retrieve useful information as per the need of the decision makers. Data mining can be used in following ways in banking sector

1. Detection of fraudulent credit card usage patterns.
2. Risk management related to attribution of loans using scorecards.
3. Find hidden correlations between different financial indicators.
4. Find hidden correlations between different financial indicators.
5. Identification of stocks trading rules from historical market data.

VIII. TASKS OF DATA MINING

A. Classification: Classification is finding models that analyze and classify a data item into several predefined classes.
B. Sequencing: Sequencing is similar to the association rule. The relationship exists over a period of time such as repeat visit to supermarket.
C. Regression: Regression is mapping a data item to a real-valued prediction variable.
D. Clustering: Clustering is identifying a finite set of categories or clusters to describe the data.
E. Dependency Modeling: Dependency Modeling (Association Rule Learning) is finding a model which describes significant dependencies between variables.
F. Deviation Detection: Deviation Detection (Anomaly Detection) is discovering the most significant changes in the data.
G. Summarization: Summarization is finding a compact description for a subset of data.

IX. DATA MINING APPLICATIONS

Data mining is a data analysis approach that has been quickly adapted and used in a large number of domains that were already using statistics. The applications areas of data mining are:

A. Medical / Pharmacy

1. Computer Assisted Diagnosis (expert systems learning).
2. Characterization/prediction of patient's response to Product dosage.

3. Identification of successful medical therapies (Successful prescription patterns).
4. Study of relations between dosage and potentially related adverse events.

B. Insurance and Health Care

1. Discovery of medical procedures that are claimed together through claims analysis
2. Identification of customers that are potential buyers for new policies.
3. Detection of behavior patterns capable of identifying risky customers.

C. Detection of Banking / Finance

1. Detection of fraudulent credit card usage patterns.
2. Risk management related to attribution of loans using scorecards.
3. Find hidden correlations between different financial indicators.
4. Identification of stocks trading rules from historical market data *Retail / Marketing* Discovery of buying behavior patterns
5. Detection of associations among customer characteristics.
6. Prediction of the probability that clients answer to mailing.

X. CONCLUSIONS

Data mining is to discover or extract knowledge or data from large amount of database. In this paper, we introduced briefly reviewed concept of data mining, issues of data mining and areas of data mining where used today. It would be helpful to researchers to focus on the various issues and challenges of data mining. 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. From the last decades, data mining and knowledge discovery applications have important significance in decision making and it has become an essential component in various organizations and fields. The complexity of data mining must be hidden from end-users before it will take the true center stage in an organization. Business use cases can be designed, with tight constrains, around data mining algorithms. Due to the enormous success of various application areas of data mining, the field of data mining has been establishing itself as the major discipline of computer science and has shown interest potential for the future developments.

REFERENCES

[1] Osmar R. Zaiane, “Principles of Knowledge Discovery in Databases”, CMPUT690, University of Alberia.
[2] Han and M. Kamber. Data Mining, “Concepts and Techniques”, Morgan Kaufmann, 2000.
[3] Bharati M. Ramageri, “Data Mining Techniques and Applications”, Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305
[4] Alex Berson, Stephen Smith, and Kurt Thearling, “Building Data Mining Applications for CRM”.
[5] Barbara, Daniel; Jajodia, Sushil (Eds.), “Applications of Data Mining in Computer Security”
[6] Gary M. Weiss, “Data Mining in the Telecommunications Industry”, Fordham University, USA.
[7] Khalid Raza, “Application of Data Mining in Bioinformatics”, “Indian Journal of Computer Science and Engineering”, .Vol 1 No 2, 114-118
[8] R.K. Mittal, Rajeev Kumar, “E-CRM In Indian Banks- An Overview”, Delhi, Business Review

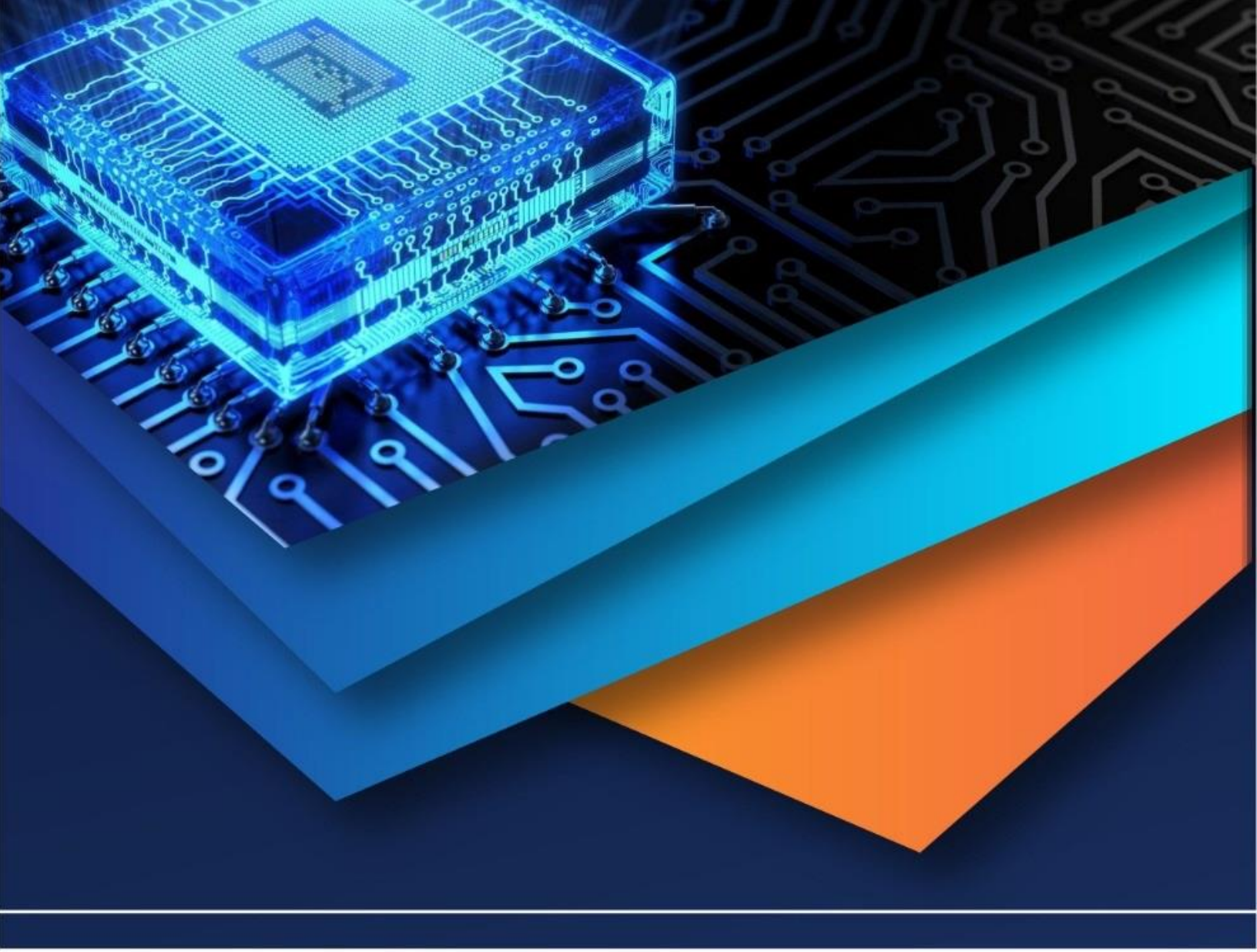
BIBLIOGRAPHY

Ms. S. Lavanya Reddy received her M.Tech (CSE) from Jawaharlal Nehru Technological University (JNTU) Hyderabad, Telangana**.** Her interested area is Big Data and Data Mining. & Data Warehousing .Presently she working as an Associate Professor in CSE Department at **Sree** Dattha Institute of Engineering & Science, Hyderabad, Telangana.

Mr. A V S M Adiseshu received his M.Tech (CSE) from Acharya Nagarjuna University, Andhra Pradesh**.** His interested area is Big Data and Data Mining. Presently he working as an Assistant Professor in CSE Department at **Sree** Dattha Institute of Engineering & Science, Hyderabad, Telangana.

Ms. P Hasitha Reddy received her M.Tech (CSE) from Jawaharlal Nehru Technological University (JNTU) Hyderabad, Telangana**.** Her interested area is Big Data, Mobile Computing and Human Computer Interaction. Presently she working as an Assistant Professor in CSE Department at Sree Dattha Institute of Engineering & Science, Hyderabad, Telangana.

Ms. Roshini K, received her M.Tech (CSE) from Jawaharlal Nehru Technological University (JNTU) Hyderabad, Telangana**.** Her interested area is Data Mining and Computer Networks. Presently she working as an Assistant Professor in CSE Department at Sree Dattha Institute of Engineering & Science, Hyderabad, Telangana.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY