International Journal of Research in Engineering, Science and Management
Volume-1, Issue-11, November-2018
www.ijresm.com | ISSN (Online): 2581-5792

Data Mining and Big Data Analytics

Jyothish Abraham[1], C. D. Aparnamol[2], Donamol Thomas[3]
[1,2,3]*Assistant Professor, Department of Computer Science, Christ College, Idukki, India*

***Abstract*: The main goal of this paper is to provide basic knowledge of data mining and big data analytics. Data mining is the semi-automatic discovery of patterns, associations, changes, anomalies and statistically significant structures and events in data. The focus will be on importance of data mining and it`s applications. Big data analytics has a vital role in the present scenario. Big data means, large data sets that may be analyzed computationally to reveal patterns, trends and associations, especially relating to human behavior and interactions. IT industry is point towards managing and maintain big data.**

***Keywords*: Data mining, stages, importance of data mining, big data, big data analytics and methodologies**

1. What is data mining

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems. The goal of data mining is to extract information from a data set and transform the information into some useful format. Data mining is the analysis step of Knowledge Discovery in Databases or KDD. The actual data mining task is semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records, unusual records and dependencies. In data mining the data are analyzed in various perspectives then classify the data and grouping the data and summarizing the identified relationships. Then it can be used for various analyzing purposes.

A. Stages in data mining

Data mining consists of following steps.

- *Data learning*: Remove noise and inconsistent data
- *Data integration*: Where multiple data sources may be combined
- *Data selection*: Where data relevant to the analysis task retrieved from database
- *Data transformation*: Where data are transformed or consolidated into forms appropriate for mining by performing summary/aggregation operations.
- *Data mining*: An essential process where intelligent methods are applied in order to extract data patterns.
- *Pattern evaluation*: To identify the truly interesting patterns representing knowledge based on some
- *Knowledge presentation*: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

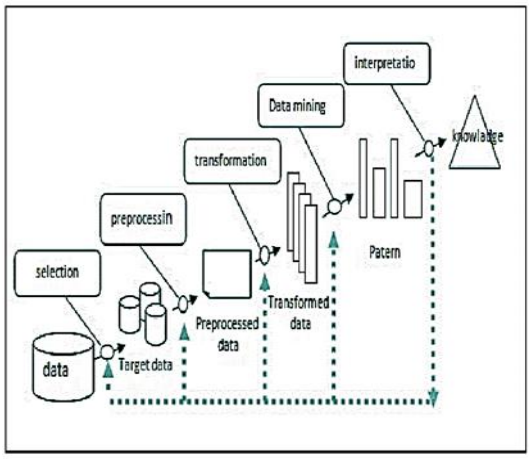

Fig. 1. Showing stages in data mining

B. Importance of data mining in present scenario

Data mining is important because, it`s concept is vast. It is used in various fields by different organizations. Five top facts why data mining and using best data mining service to mine it is important.

- Millions terabytes of data is getting generated and companies are processing it using big data technology.
- The business organizations and many industry use different data mining technologies to collect information in different aspects.
- Another point of view is that you require data from website. For acquiring data from websites data mining or web mining is needed.
- Security is an important aspects of every organization. Data mining has an important role in tracking of records and data with domain knowledge.
- Nowadays, e-commerce has a vital role in our daily life. So, data mining also play an important role in a human life.

C. Financial data analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows –

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.

- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

D. Retail industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web. Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry –

- Design and Construction of data warehouses based on the benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

E. Telecommunication industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business. Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services –

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

F. Biological data analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

G. Other scientific applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

H. Intrusion detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.
- Products and promotions to appeal to specific customer segments.

I. Future healthcare

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

J. Market basket analysis

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behavior of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

K. Education

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behavior, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

L. Manufacturing engineering

Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

M. CRM

Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyses the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

N. Fraud detection

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

O. Lie detection

Apprehending a criminal is easy whereas bringing out the truth from him is difficult. Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists. This filed includes text mining also. This process seeks to find meaningful patterns in data which is usually unstructured text. The data sample collected from previous investigations are compared and a model for lie detection is created. With this model processes can be created according to the necessity.

P. Customer Segmentation

Traditional market research may help us to segment customers but data mining goes in deep and increases market effectiveness. Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the customers. Market is always about retaining the customers. Data mining allows to find a segment of customers based on vulnerability and the business could offer them with special offers and enhance satisfaction.

Q. Financial marketing

With computerized banking everywhere huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers may find these information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

R. Corporate surveillance

Corporate surveillance is the monitoring of a person or group's behavior by a corporation. The data collected is most often used for marketing purposes or sold to other corporations, but is also regularly shared with government agencies. It can be used by the business to tailor their products desirable by their customers. The data can be used for direct marketing purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails.

S. Research analysis

History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualization and visual data mining provide us with a clear view of the data.

T. Criminal investigation

Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Text based crime reports can be converted into word processing files. These information can be used to perform crime matching process.

U. Bio informatics

Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

2. Big data analysis

Big data may be the next big thing in the IT world. Big data burst upon the scene in the first decade of the 21st century. Firms like Google, eBay, LinkedIn and Facebook were built around big data from the beginning.

A. What Is Big Data?

The term big data was first used to refer to increasing data volumes in the mid-1990s. In 2001, Doug Laney, then an analyst at consultancy Meta Group Inc., expanded the notion of big data to also include increases in the variety of data being generated by organizations and the velocity at which that data was being created and updated. Big data is similar to 'small data' but bigger in size. Big data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

B. Characteristics of Big Data –V3s

1) Volume

Volume refers to the amount of data generated through websites, portals and online applications. Volume encompasses the available data that are out needed to be accessed for relevance. For example, Facebook has 2 billion users, YouTube 1 billion users etc. You can now imagine the instantly large amount or volume of data is generated every minute and every hour.

2) Velocity

Velocity refer to the speed with which data are being generated. High frequency stock trading reflect market changes within microseconds. Machine to machine processes exchange data between billions of devices every day. Example, everyday 9000 million photos are uploaded on Facebook.

3) Variety

Variety in Big Data refers to all the structured and unstructured data that has the possibility of getting generated either by humans or by machines. The most commonly added data are structured -texts, tweets, pictures & videos. However, unstructured data like emails, voicemails, hand-written text, ECG reading, audio recordings etc., are also important elements under Variety. Variety is all about the ability to classify the incoming data into various categories.

C. Big data analytics

The volume of data that one has to deal has exploded to unimaginable levels in the past decade, and at the same time, the price of data storage has systematically reduced. Private companies and research institutions capture terabytes of data about their users' interactions, business, social media, and also sensors from devices such as mobile phones and automobiles. The challenge of this era is to make sense of this sea of data. This is where big data analytics comes into picture. Big Data Analytics largely involves collecting data from different sources, mange it in a way that it becomes available to be consumed by analysts and finally deliver data products useful to the organization business. The process of converting large amounts of unstructured raw data, retrieved from different sources to a data product useful for organizations forms the core of Big Data Analytics.

D. Traditional data mining life cycle

In order to provide a framework to organize the work needed by an organization and deliver clear insights from Big Data, it's useful to think of it as a cycle with different stages. It is by no means linear, meaning all the stages are related with each other. This cycle has superficial similarities with the more traditional data mining cycle as described in CRISP methodology.

1) CRISP-DM methodology

The CRISP-DM methodology that stands for Cross Industry Standard Process for Data Mining, is a cycle that describes commonly used approaches that data mining experts use to tackle problems in traditional BIG data mining. It is still being used in traditional BIG data mining teams. CRISP-DM was conceived in 1996 and the next year, it got underway as a European Union project under the ESPRIT funding initiative. The project was led by five companies: SPSS, Teradata, Daimler AG, NCR Corporation, and OHRA (an insurance company). The project was finally incorporated into SPSS. The methodology is extremely detailed oriented in how a data mining project should be specified.

Let us now learn a little more on each of the stages involved in the CRISP-DM life cycle −

- *Business Understanding* − This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition. A preliminary plan is designed to achieve

the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used.

- *Data Understanding* − The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.
- *Data Preparation* − The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.
- *Modeling* − In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, it is often required to step back to the data preparation phase.
- *Evaluation* − At this stage in the project, you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model thoroughly and review the steps executed to construct the model, to be certain it properly achieves the business objectives.

A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment − Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process. In many cases, it will be the customer, not the data analyst, who will carry out the deployment steps. Even if the analyst deploys the model, it is important for the customer to understand upfront the actions which will need to be carried out in order to actually make use of the created models.

2) SEMMA methodology

SEMMA is another methodology developed by SAS for data mining modeling. It stands for Sample, Explore, Modify, Model, and Asses. Here is a brief description of its stages −

- *Sample* − The process starts with data sampling, e.g., selecting the dataset for modeling. The dataset should be large enough to contain sufficient information to retrieve, yet small enough to be used efficiently. This phase also deals with data partitioning.
- *Explore* − This phase covers the understanding of the data by discovering anticipated and unanticipated relationships between the variables, and also abnormalities, with the help of data visualization.
- *Modify* − The Modify phase contains methods to select, create and transform variables in preparation for data modeling.
- *Model* − In the Model phase, the focus is on applying various modeling (data mining) techniques on the prepared variables in order to create models that possibly provide the desired outcome.
- *Assess* − The evaluation of the modeling results shows the reliability and usefulness of the created models.

The main difference between CRISM–DM and SEMMA is that SEMMA focuses on the modeling aspect, whereas CRISP-DM gives more importance to stages of the cycle prior to modeling such as understanding the business problem to be solved, understanding and preprocessing the data to be used as input, for example, machine learning algorithms.

3. Conclusion

Data mining is a field of intersection of computer science and statistics used to discover patterns in the information bank. The main aim of the data mining process is to extract the useful information from the dossier of data and mold it into an understandable structure for future use. There are different process and techniques used to carry out data mining successfully. Nowadays, all era needs data mining for their successful running. Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful information for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible.

References

[1] Jaawei Han and Micheline Kamber, “Introduction To Data Mining” in Data Mining Concepts and Techniques, second Edition, Morgan kaufmann publishers, 2006, pp. 5-9.
[2] Limsoon Wong, “Data Mining Techniques” Acm Sigmod Record (Sigmod Rec), pp. 2-6, August 2003.
[3] Margaret Rouse, “Enterprise guide to big data in cloud computing”.