International Journal of Research in Advent Technology, Vol.7, No.5S, May 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

A Survey on Data Mining: Tools, Techniques, Applications and Major Issues

Ms. Sushma Malik
Assistant Professor(IT)
IINTM, Janakpri, New Delhi

Abstract— In the IT era, information plays a fundamental role in every phase of human life. For every stakeholder, it is very essential to accumulate the data from the various data sources and generate useful information and knowledge from that stored data. Due to vast usage of electronic devices like computers and smart phones, there is explosive growth in data collection. For to analyze that huge amount of data and dig out the meaningful information and to find the conclusion, user needs the special kind of tools called data mining. Data mining is the process though which user finds useful, interesting, and formerly unknown patterns from massive data. It is a powerful technology that helps companies to retrieve important and useful information from their data warehouses. It supports decision making process by providing quick and accurate information. Data mining also analyze the data to identify relationship between different data elements or entities. The main aim of this research paper is to provide a review on various data mining techniques and tools. This paper also includes the formal review of data mining applications in various fields.

Keywords— Data mining (DM), (KDP) Knowledge discovery process, classification, Mining Tools, data ware house**.**

1. INTRODUCTION

There has been a remarkable amplify data which is stored in binary format since last few decades. The database size has been increased day by day. The data may be in different format such as simple with numerical and text and in more complex format with spatial data, audio, video and hypertext documents from the different sources. Large amount of data is required to engender the meaningful information [1][2]. Data can also be in structured and unstructured format. The retrieval of data from the database is not enough but requires a tool to extract the meaningful information, change that data into summarization form data and also analyze the data to identify relationship between different data elements or entities. The only answer to all above is 'Data Mining'. DM is the logical process that is used to dig out the analytical information from big size of data bases or warehouses. The main aim of DM is to dig out the various patterns of data that were not earlier identified [3].

DM is also identified with other names like information discover, knowledge extraction and data pattern processing [4]. DM is just one step in the Knowledge discovery process, but some researchers using it as a synonym for KDP. In the knowledge discovery process, all the steps like cleaning of data, integration of data which is collected from the various sources, selection of required data, data transformation, DM and knowledge representation, can execute one after the other to find the useful information from the database.[1]

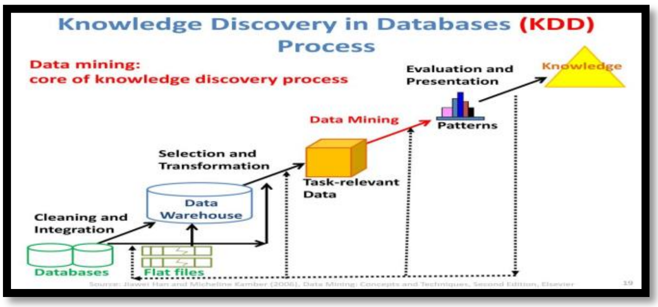
Knowledge Discovery in Databases (KDD) Process
Data mining:
core of knowledge discovery process
Evaluation and Presentation
Knowledge
Data Mining
Patterns
Selection and Transformation
Task-relevant Data
Data Warehouse
Cleaning and Integration
Databases
Flat files
Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier
19

Figure1: Knowledge Discovery Process

Types of data that can be mined

1. **Flat files:** Flat files are the simplest files which contain data in text or in binary format. These files are the main sources for DM process. These files contain the transactional data, time-series data, scientific data and many more.
2. **Relational Databases:** It is collection of rows and columns which are called table, where columns represent attributes and rows represent tuples.
3. **Data Warehouses:** A data warehouse act as the repository of data collected from multiple data sources. Data warehouses provide the opportunity to analyze data which are collect from the different sources under the same roof.

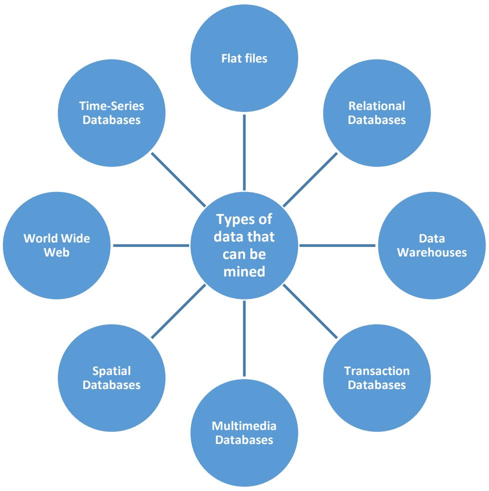

Figure 2: Types of data that can be mined

4. **Transaction Databases:** This kind of database contains a set of records that representing transactions with a time stamp and an identifier.

5. **Multimedia Databases:** It contains video, images, and audio and text files. Object-oriented databases basically stored this kind of data. The higher dimensionality of the multimedia files makes DM process more challenging.
6. **Spatial Databases:** That kind of databases is basically the collection of geographical information of places like maps, and global or regional positioning. Such kind of data is creating new challenges for the DM algorithms.
7. **World Wide Web:** WWW contains the heterogeneous data with dynamic repository which change day to day. Data in the WWW is organized in hyperlinked documents like text, video, audio, raw data, and even applications.
8. **Time-Series Databases:** This type of database to be full of time related data such stock exchange data or logged activities in the various applications. These type of databases are continues increase in size due to adding of new data which sometimes increase real time analysis challenging for the researcher.[1]

2. DATA MINING IMPLEMENTATION PROCESS

DM implementation process consists of six phases of the life cycle. The sequence of the phases is not fixed. Moving between different phases is always depends on the outcome of each phase. The main phases are:

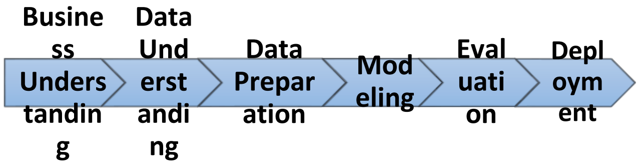

Figure 3: Data Mining Implementation Process

1. **Business Understanding:** The main focus of this phase is to recognize the objectives and requirements of the projects on the basis of business perspective and then create DM goals to achieve the business objectives.
2. **Data Understanding:** This phase start after the collection of data. After collecting data, user gets the familiarity with the data and identifies data quality problems. And after that user identify the interesting data subset to design hypotheses for hidden information.
3. **Data Preparation:** the main function of this phase is cleaned, constructed and formatted the data in the desired format which are collected from the various sources.
4. **Modeling:** On the basis of business objective, a suitable modeling technique should be applied on the prepared dataset and results are assessed by the user to identify that model fulfill the DM objectives.
5. **Evaluation:** In evaluation phase, the result of modeling techniques are evaluated and reviewed with the business objectives in the first phase. The new objectives of business may be raised in this phase because the new patterns have been discovered by the modeling result.
6. **Deployment:** The function of this phase is the information discovered during the DM should be made easy to understand for non technical users. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable DM process across the enterprise.[1]

3. DATA MINING TECHNIQUES

DM is a method through which relevant information is extract from unstructured data. The DM can be implemented with:

Descriptive model: This model presents the extracted information in concise form.

Predictive model: through this model, data miner predicts the unidentified value for a specific variable [4].

1. **Classification:** The most useful DM technique is classification. This technique is used to predicting the class for new data instances. This technique is based on the supervised learning which means the desired output for a given input is known. By providing training the data can be trained and we can predict the future of data means predict the class to which data can belong.

The classification process of DM involves two steps:

- Learning and
- Classification.

In the first learning step, the data sets are evaluate with by classification algorithm and in second classification step, the rules of classification are implemented or apply on test data to estimate the accuracy of the classification rules. The rules can be applied to new data, if the accuracy of the rule is acceptable. This technique is basically used in credit risk and fraud detection applications areas. [1][5]

2. **Regression:** It is another DM technique which is based on supervised learning. It is used for the prediction of the result from the available data sets. Target value is known in the regression techniques. For example, the behavior of the child can be easily predicted on family history. It estimates the value by comparing already known value with predicted values.[5]
3. **Time Series Analysis:** Time series analysis using statistical techniques to explain a time-dependent series of data. This technique is using to generate the future predictions (forecasts) based on already known past events [9]. For example stock market predict the value of shares in the market on the basis of past data.[4]
4. **Prediction:** It is a DM technique which is used to find out relationship between the different variables and these variables may be independent or dependent. For example the prediction method is used to predict the profit(dependent variable) on the past sale(independent variable). So the profit prediction can be done by using the past sale and profit data.[1]

5. **Clustering:** This technique is based on the unsupervised learning means the desired output for a given input is not known. It is used to makes meaningful cluster of object which has similar characteristics. In this technique classes are define for the objects and objects are placed in each class, but in the classification techniques, objects are disperse into the predefined classes. For example, image processing and pattern recognition uses the clustering technique. Clustering can be categorized in two form as follows:
   Hard clustering: In the hard clustering, similar object can be belonging to single group.
   Soft clustering: In the soft clustering, similar object can be belonging to different groups. [5][1][4]
6. **Summarization:** Summarization technique of DM is applied to abstract the data. It provides the overview of data. For illustration, the race can be precise in total minutes and seconds need to cover that distance.[4]
7. **Association Rule:** Association is one of the best known DM techniques. In this, a new pattern is discovered on the basis of association between the different items in the same transaction. Market Basket Analysis to classify the set of items that buyers purchase together is used association rule on purchase data. [1][4]
8. **Sequence Discovery:** Sequence Discovery method or technique is applied on database to uncover the relationship between the data. For example, it is used in scientific experiment, analysis of DNA sequence, and natural disaster.[4]
9. **Text Mining**: Text mining method implement on text data which include documents, emails, messages, and html files. Text mining can be classified as document processing, document summarization, indexing, topic clustering and mapping. It is normally used in educational area and business area to analyze the text data. Each and every Organization has huge amount of document collection and to text mining is implement on that data to retrieve the useful and interesting information. [6].
10. **Decision trees:** Decision tree is mostly used because user can understand its functioning easily. The roots of the decision tree is a simple condition with multiple answers and each answer further then direct to a number of conditions that assist the user to build the final decision. It is effective tool in number of areas like text mining, machine learning pattern recognition.[1]

4. A BRIEF OVER VIEW OF DATA MINING TOOLS

For performing the analysis of data by using any DM technique, required the knowledge of different tools. Numbers of DM tools are present in the market. On the basis of functioning DM tools can be classified as:

1. **Traditional Data Mining Tools:** Traditional DM tools works with existing databases stored on organization servers. These tools understand the stored data by using pre-defined algorithms and queries written out in a database specific programming language.
2. **Dashboards**: Dashboard is software that present on an end-user's desktop or tablet and produce the real time fluctuations reports when data move into the database. It is used by managers and other users to track the effect of events and other influences on data streams in real time.
3. **Text-mining Tools:** The text mining tool mine data from the different kind of text like word document, text files and PDF files. It scans the content and converts into formats that is well suited with the database and help the user to access the data with an easy and convenient way without to open the different applications. [1]

Some of the popular DM tools are as follows:

Tool 1-Weka: Weka means Waikato Environment for Knowledge Analysis. Weka is a Java based open source tool of DM and is a set of many machine learning algorithms for DM tasks like pre-processing on data, classification, clustering, and association rule extraction. These algorithms can either be applied directly to a data set or can be called from user Java code. [1]
Tool 2- KEEL: Knowledge Extraction based on Evolutionary Learning is an application package of machine learning software tools. It is designed to provide solution of DM problems and also assessing evolutionary algorithms. It has a vast collection of libraries for pre-processing and post-processing for data manipulating in knowledge extraction [1].

Tool 3- R: Revolution(R) is free programming language software and it provide environment for arithmetical computing and graphics. It is the most widely used software by mathematical and data miners for developing statistical software and data analysis process. The biggest advantages of R tool that it is an entirely open sourced which means that it can be downloaded by user very easily and is available at free of cost. Another best part of it is that any user is allowed to enhance its code and add new packages on the basis of requirement. [11]

Tool 4- KNIME: Konstanz Information Miner is an open source for data analytics. It is a powerful tool with good GUI and based on Eclipse platform with easily extensible. It is popular amongst the financial data analysts. [1]

Tool 4- RAPIDMINER: It is a software platform developed by the company of the same name. It helps enterprises in predictive analysis of their business process with its user friendly environment, library and machine learning algorithms with its all-in-one platform like Rapid Miner Studio. It is also used in research and education area. It is based on client-server architecture. [1][7]

Tool 5- ORANGE: Orange is a Python based powerful and open source tool for both new users and for experts. Users can focus on the data analysis instead on programming part of it. [7]

Tool 6-UIMA: UIMA stands for Unstructured Information Management Application and it help to analyze the large amount of unstructured data to information which is useful for the end users. It enables application to be decomposed into components. Working of framework is to manage these components and flow between them. The main goal of UIMA is to transform unstructured information to structured information and thus to build the bridge between the unstructured and the structured world. [7][12]

5. DATA MINING APPLICATION

Number of field adapted DM technology because of its fast access of data and extract the meaning from a large amount of data. Some of them are listed below:

Data Mining in Education Sector: We are applying DM in education sector then new emerging field called "Education Data Mining". It is blooming field which provides knowledge from educational Environment data. The main aim of EDM is recognized as predicting students' learning behavior, emotions and skills of the students. The student's data is used to understand the student learning behavior to predict their results. With this kind of study, the educating methods can be improved by taking the accurate decision at accurate time. [4][8]

Data Mining in Market Basket Analysis: Market Basket Analysis based on shopping database. It is using the association rule of mining to find the goods that customers are frequently purchase together and understand the purchasing behavior of the customer. The sellers can use this information in their store by putting these products in close nearness of each other and making them more visible to customers and accessible for them at the time of shopping [4][8][9]

Data Mining in Banking and Finance: DM has been extensively used by the banking and financial markets. In the banking area, DM is used to detect credit card fraud, estimate risk level on loans. In the financial markets, DM technique used for price prediction, stock forecasting and so on by using the neural networks DM technique. [4][9]

Data Mining in Telecommunication: Telecommunications area using the DM because it has huge amount of customer's data and highly competitive environment in it. Telecommunication companies improve their marketing efforts and better management of telecommunication networks by using the DM techniques. [4]

Data Mining in Agriculture: DM is an emerging technology in farming field for crop yield analysis with respect to four parameters like year, rainfall, production and area of sowing. The main problem of agricultural problem is yield prediction which is solved based on the available data employing DM techniques.[4]

Data Mining in Bioinformatics: Bioinformatics is the biological field that deals with the gathering, storage and understands the biological data with the DM tool and help the researchers to develop a better understanding of biological mechanism to discover new treatments for dieses and knowledge of life. [4][1]

Data mining in Governments: Government agency can digging the database of the data and analyzing the financial transactions records of the users to design the models that help to detect money laundering or criminal activity. [9]

6. MAJOR ISSUES IN DATA MINING

DM is not a simple assignment because the algorithms used for it, can be very complex and data is some time scattered so not always accessible at one place. For analysis of data, it should be integrated from various heterogeneous data sources. The major issues regarding the analysis of data in DM are [13]:

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

Mining Methodology and User Interaction

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases**: It is necessary for DM to cover broad range of knowledge discovery task because different users may be interested in different kinds of information.
- **Interactive mining of knowledge at multiple levels of abstraction**: The DM process needs to be interactive because it allows users to refining DM requests based on the returned results.
- **Presentation and visualization of data mining results**: The DM results need to be easily understandable by the users and visual representative.
- **Handling noisy or incomplete data**: Data should be integrated from the various sources in the database. The data cleaning methods are required to handle the noise and incomplete data while mining the database. If the data cleaning methods are not implemented there then the accuracy of the discovered patterns will be poor and misguide the user.
- **Pattern evaluation**: A DM system can discovered number of patterns. Many of discovered patterns are uninteresting to the user, representing common knowledge or lacking novelty.

Performance Issues

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms**: DM algorithm must be efficient and scalable to effectively extract the information from huge amount of data in databases.
- **Parallel, distributed, and incremental mining algorithms**: The huge size of databases,

complexity of some DM methods and wide distribution of data are some motivating factors for the development of parallel and distributed DM algorithms. These algorithms divide the data into partitions which is further processed in a parallel manner and then the results from the partitions are merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data**: The database may contain complex data objects like audio, video and unstructured data. So it's not possible for one system to mine all these type of data. To fulfill this requirement, a specific DM systems should be constructed for mining specific kinds of data.
- **Mining information from heterogeneous databases and global information systems:** The data is collected from different data sources on LAN or WAN in the database and some time these data source may be structured, semi structured or unstructured. Therefore mining the knowledge from that kind of databases are always adds challenges to DM.

7. CONCLUSION

Data mining is an application oriented technology. It is not only a simple search and transfer of data from the particular database but also analyze and integration of data from the different data bases. It helps to find the solutions of the practical problems and also help to find the relation between different entities attributes and even help to predict the future activities by using the existing data.

This paper presents the general idea of DM, its techniques and tools and its application in various areas. DM techniques helps in finding the different patterns to the user to decide the future development in businesses to need to improve. It play vital role in every area where the data is generated that's and that's why it is considered one of the most important tool in database or data ware house. In future work we review various tools and their significance and also comparing their functioning.

REFERENCES

[1] Vikas Gupta, Prof. Devanand, "A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues", International Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013, ISSN 2229-5518.

[2]S. P. Deshpande, V. M. Thakare, "Data Mining System And Applications: A Review", International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010. DOI: 10.5121/ijdps.2010.1103

[3] Parminder Kaur, Qamar Parvez Rana, "Comparison of Various Tools for Data Mining", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 3 Issue 10, October- 2014.

[4] Smita, Priti Sharma, "Use of Data Mining in Various Field: A Survey Paper", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. V (May-Jun. 2014), PP 18-21 www.iosrjournals.org.

[5] Mansi Gera, Shivani Goel, "Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity", International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 18, March 2015

[6] Hussain Ahmad Madni, Zahid Anwar, Munam Ali Shah, "Data Mining Techniques and Applications – A Decade Review", Conference Paper · September 2017 DOI: 10.23919/IConAC.2017.8082090.

[7] Deepashri.K.S, Ashwini Kamath, "Survey on Techniques of Data Mining and its Application", International Journal of Emerging Research in Management & Technology, ISSN: 2278-9359(volume-6, Issue-2), February2017.

[8] Hemlata Sahu, Shalini Shrma, Seema Gondhalakar, "A Brief Overview on Data Mining Survey", International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3.

[9] Maqsud S. Kukasvadiya, Nidhi H. Divecha, "Analysis of Data Using Data Mining tool Orange" 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939.

[10] Luís C. Borges, Viriato M. Marques and Jorge Bernardino, "Comparison of Data Mining techniques and tools for data classification" Conference Paper · July 2013 DOI: 10.1145/2494444.2494451.

[11] N.Deepika, "A Novel Survey on Different Mining Tools", IJCSMC, Vol. 5, Issue. 1, January 2016, pg.233 – 240, ISSN 2320–088X.

[12] https://uima.apache.org/doc-uima-why.html.

[13]https://www.tutorialspoint.com/data_mining/dm_issues.html