International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359

Research Article **February 2013**

Data Mining Tools and Trends – An Overview

S.Hameetha Begum[*]
Senior Lecturer, Dept of Computing,
Muscat College, Oman

Abstract—

Today Information Technology plays a vital role in every aspects of the human life. It is very essential to gather data from different sources. This data can be stored and maintained to generate information and knowledge. This information and knowledge has to be disseminated to every stake holders for the effective decision making process. Due to the increase in the data, it is important to extract knowledge/information from the large data repositories. Hence, Data mining has become an essential factor in various fields including business, education, health care, finance, scientific etc., To analyse this vast amount of data and depict the fruitful conclusions and inferences, it needs specific data mining tools such as R, Weka, Orange. This paper discusses the knowledge discovery process, data mining, various open source tools and improvements in the field of data mining from past to the present and explores the future trends.

Keywords – Data Mining, Knowledge Discovery Process, Data Mining Tools, Weka, R, Orange. Data Mining Trends,

I. Introduction

The development of Information technology has paved way to generate large amount of databases and huge data in various areas. The research in databases and information technology has given rise to approach to store and manipulate precious data for further decision making [1]. Data mining is a process to extract the implicit information and knowledge by extracting from the mass, incomplete, noisy, fuzzy and random data with knowing the data well in advance and which is potentially useful to various fields [2]. This paper is organized as follows. Section II describes knowledge discovery process Section III explains Data mining definition, advantage, disadvantage and its challenges Section IV demonstrates categories of data mining tools. Section V explains the open source tools for data mining like R, Weka, Orange, RapidMiner and Tanagara. Section VI focuses on the trends of data mining and final section is the conclusions.

II. Knowledge Discovery Process

The process of discovering useful knowledge from a huge data is called as Knowledge Discovery in Database (KDD) and which is often referred to as Data mining. While data mining and knowledge discovery in databases are normally treated as synonyms, but, in fact data mining is a part of knowledge discovery process. The KDD process comprises of few steps as shown in Fig. 1and explained as follows[2]:

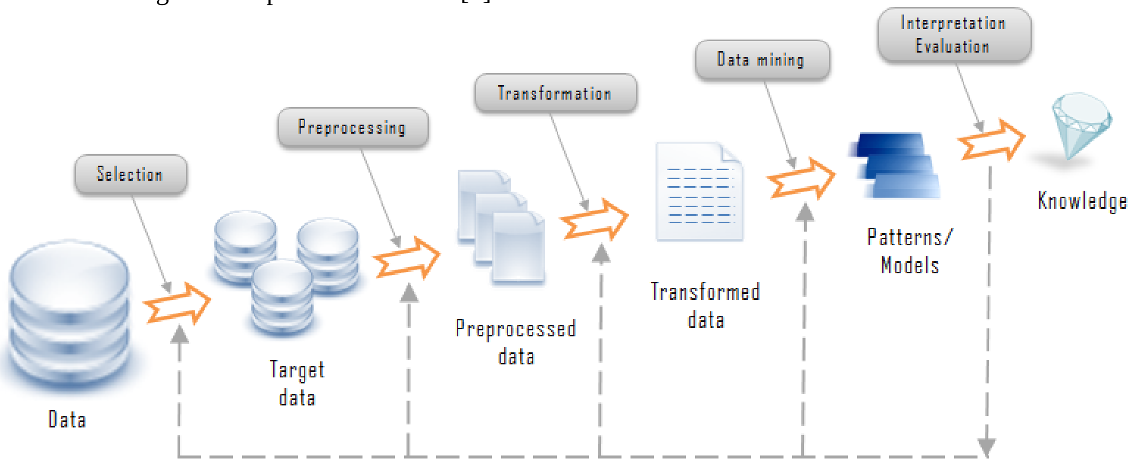

Fig. 1 Knowledge Discovery Process [3]

Data collected from multiple sources often heterogeneous is integrated into a single data storage called as target data. Data relevant to the analysis is decided on and retrieved from the data collection. Then, it is pre-processed and transformed into an appropriate standard format. Data mining is a crucial step in which intelligent algorithm/techniques are applied to extract meaningful pattern or rules. Finally, those patterns and rules are interpreted to new or useful knowledge or information [1][3].

III. Data Mining

A. Definition of Data Mining

"Data Mining represents a process developed to examine large amounts of data routinely collected. The term also refers to a collection of tools used to perform the process. Data collected from various areas such as marketing, health, communication, etc., are used in data mining." [4]

"Data Mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organisations focus on the most important information in their data warehouse."[5]

Questions those traditionally were too time consuming to resolve can be answered by the data mining tools in an effective manner. This helps to find the hidden patterns, predictive information that facilitates the experts with solution outside their expectations [5]. The goal of data mining is to extract knowledge from dataset in human-understandable structures. In recent years data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and engineering [6].

B. Advantages of Data Mining in various applications

Advantages of using data mining in various applications such as Banking, Manufacturing and production, marketing, health care etc., are as follows[7]:

1) *Banking:* Data mining supports banking sector in the process of searching a large database to discover previously unknown patterns; automate the process of finding predictive information. Data mining helps to forecast levels of bad loans and fraudulent credit cards use, predicting credit card spending by new customers and predicting the kinds of customer best respond to new loan offered by the backs.
2) *Manufacturing and production:* Data mining helps to predict the machine failures and finding key factors that control optimization of manufacturing capacity.
3) *Marketing:* Data mining facilitates marketing sector by classifying customer demographic that can be used to predict which customer will respond to a mailing or buy a particular product and it is very much helpful in growth of business.
4) *Health-Care:* Data mining supports a lot in health care sector. It supports health care sector by correlating demographics of patients with critical illnesses, developing better insights on symptoms and their causes and learning how to provide proper treatments
5) *Insurance:* Data mining assist insurance sector in predicting fraudulent claims and medical coverage cost, classifying the important factors that affect medical coverage and predicting the customers' pattern which customer will buy new policies.
6) *Law:* Law enforcement is helped by data mining by monitoring the behaviour patterns of the criminals. Tracking crime pattern, locations and criminal behaviours, identifying various attributes to data mining, assist in solving criminal cases.
7) *Government and Defence*: Data mining helps to forecast the cost of moving military equipment and predicting resource consumption. Apart from that it assists in testing strategies for potential military engagements and improving homeland security by mining data from many sources.
8) *Brokerage and Securities trading:* Data mining assists in predicting the change in bond prices and forecasting the range of stock fluctuation determining when to buy or sell stocks.
9) *Computer hardware and software:* Predicting disk-failures and potential security violations can be done by data mining.
10) *Airlines:* It supports in checking the feasibility of adding routes to increase the business profit and to decrease the loss by capturing data on where passengers are flying and the ultimate destination of passengers.

C. Disadvantages of Data Mining

The disadvantages of data mining are explained as follows:

1) *Privacy Issues*

One of the disadvantages is a personal privacy issue. In recent years, with the boom of internet, the concerns about privacy have increased tremendously. Because of this privacy concern, individuals like internet users, employees, customers are afraid that unknown person may have access to their personal information and then use that information in an unethical way and this may cause harm to them. Although, several laws have protected the users to sell or trade personal information between different organisation, selling personal information have occurred [2][7].

2) *Security Issues*

Another biggest disadvantage is security issue which is always a major concern in information technology. Companies have a lot of personal information about the employees and customers including social security number, birthdates, payroll etc., and it is also available in online. But, they do not have sufficient security systems in place to protect this information. They have been a lot of cases where hackers access and stole personal data of customers [2][7].

3) *Misuse of Information/Inaccurate information*

Trends obtain from the data mining intended to be used for business or some ethical purpose. However it may be misused for other unethical purpose. Unethical businesses or Individual may use the information to take advantage of vulnerable people or to discriminate against a certain group of people. Apart from that, data mining techniques is not cent percent accurate one. Thus mistakes may happen which can have serious consequence [8].

D) *Challenges of Data Mining*

There are many challenges faced by the data mining and these challenges of data mining are pointed as follows[2][9][10]:

- Scalability
- Complex and Heterogeneous Data
- Network Setting
- Data Quality
- Data Ownership and Distribution
- Dimensionality
- Privacy preservation
- Streaming Data

IV. Categories Of Data Mining Tools

Most of the data mining tools can be classified into three categories: Traditional data mining tools, dash boards and text-mining tools. Description of each is as follows [11]:

A. *Traditional Data Mining Tools*

Traditional mining programs help the companies to establish data patterns and trends by using various complex algorithms and techniques. Some of these tools are installed on the desktop computers to monitor the data and emphasize trends and others capture information residing outside a data base. Majority of these programs are supported by windows and UNIX versions. However, some software specializes in one operating system only. In addition to that some may work in only one database type. But, Most of the software will be able to handle any data using online analytical processing or a similar technology.

B. *Dashboards*

Dashboards reflect data changed and update on screen. Dashboards is normally installed in computers to monitor information in a database and it reflects data changes and updates the data in the form of a chart or table on the screen. It enables the user to see how the business is performing. Historical data can be referenced and checks against the current status in order to see the changes in the business. By this way, dashboards is very easy to use and helps the manager a lot with great appeal to have an overview of the company’s performance.

C. *Text-Mining Tools*

The third type of data mining tools is called as a text-mining tool because of its ability to mine data from different kind of text starting from Microsoft Word, Acrobat PDF documents to simple text files. This provides facility of scanning the content and converts the selected into a format that is compatible with the tools database without opening different applications.

V. Open Source Tools for Data Mining

Different types of data mining tools are available in software market, each with its own strengths and weaknesses. Some of the data mining tools available today are explained as follows:

A. *R*

R is an open source programming language and environment for statistical computing and graphics. R provides a wide variety of graphical and statistical techniques such as linear and non-linear modelling, classical statistical tests, time-series analysis, classification clustering and is highly extensible. Researchers in various fields of applied statistics have adopted R for statistical software development and data analysis. Extensibility and superb data visualisation are the two main reasons for the success of R [6][12].

B. Weka

Weka is a collection of machine learning algorithms for data mining tasks and well suited for developing new machine learning schemes. Weka is a java based software capability of working under various operating systems and contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. The algorithms can either be applied directly to a dataset or called from a user's java code. Weka is probably the most successful open source data mining software which has inspired by the development of other programs with more sophisticated graphical user interface and better visualization methods [13][14].

C. Orange

Orange is an open source data mining and visualisation software with active community and which helps novice and experts for their analysis. It has the ability to work under various platforms like windows, Mac Os C and GNU/Linux operating systems and it's packed with data analytics features. It enables design of data analysis process through user friendly visual programming or python scripting. Hence, this can be used as a scripting language for respective tasks of data mining. It represents most major algorithms for data mining and contains different visualisation, from scatter plots , bar charts, trees to dendrograms, networks and heatmaps. It remembers user's choices, suggests most frequently used combinations, and intelligently chooses which communication channels to use. It has specialised add-ons like Bioorange for bio informatics[15].

D. Rapid-I RapidMiner

Rapid-I RapidMiner is an open source system for data mining which is available as a standalone application for data analysis and as a data mining engine for the integration into own products. It has ability to runs on major platform and operating systems. It is powerful but intuitive graphical user interface for the design of analysis processes. It offers data integration, analytical ETL, Data analysis and reporting in one single suite. It provides a graphical process design for standard tasks and scripting language for arbitrary operations [16].

E. Tanagara

Tanagara is open source data analysis software for academic and research purposes which proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area [11].The main purpose of Tanagra is to provide researchers and students to use data mining software in an easy way by conforming to the present norms of the software development and allowing to analyse either real or synthetic data. The second purpose is to propose an architecture allowing the users to add to add their own data mining methods it helps to compare their performances. It acts more as an experimental platform in order to do the essential work, dispensing them to deal with the unpleasant part of the data management. Last purpose is to give the direction to novice developers in diffusing a possible methodology for building this kind of software. It can be considered as a pedagogical tool for learning programming techniques since it permits to access the source code, to look pattern of the software how it is built, the problems to avoid, key steps of the project, tools used and code libraries used for the project[17].

F. Top Ranked Data Mining Software in 2012

KD Nuggets has conducted an annual poll in the year 2012 by asking question "What analytics/data mining software you used in the past 12 months for a real project (not just evaluation)". R was the top-ranked data mining solution chosen by 30.7% of poll respondents. Microsoft Excel was second at 29.8% and RapidMiner ranked third. In this poll, R was also ranked as the most popular language for implementing data mining applications beat out SQL and java[18][19].

VI. Trends in Data Mining

A. Historical Trends

Data mining application era was perceived in early 1980s principally focused on single tasks driven by research tools. Data mining is helpful in various disciplines like Data Base Management Systems (DBMS), Artificial Intelligence (AI), Machine Learning (ML) and Statistics. Historical trends of data mining are explained as follows [4]:

1) *Data Trends:* Data mining algorithm work best with the numerical data especially collected from a single data base and various data mining techniques have developed for flat files, traditional and relational database where the data is mostly represented in the tabular form. Afterwards, with the convergence of Statistics and Machine Learning pave way to the evolution of various algorithms to mine the non numerical data and relational data bases.

2) *Computing Trends*: Development in fourth generation programming language influenced much in the field of data mining and various related computing techniques. Initially, most of the algorithms engaged to work only on statistical techniques. Various computing techniques such as AI, ML and pattern reorganization evolved to do the data mining tasks in ease manner. Various data mining techniques like Induction, Compression, approximation and other algorithms developed to mine the large volume of heterogeneous data stored in the data warehouse.

B. Current Trends

Advancement in data mining with various integrations and implications of the methods and techniques have formed the present data mining applications to tackle the various challenges. The current trends of data mining application are described as follows [4]:

TABLE I
CURRENT DATA MINING AREAS AND TECHNIQUES TO MINE THE VARIOUS DATA FORMAT [4]

Data mining type	**Application Areas**	**Data Formats**	**Data mining Techniques/Algorithms**
Hypermedia data mining	Internet and Intranet Applications.	Hyper Text Data	Classification and Clustering Techniques
Ubiquitous data mining	Applications of Mobile phones, PDA, Digital Cam etc.	Ubiquitous Data Traditional data mining techniques drawn from the Statistics and Machine Learning	Traditional data mining techniques drawn from the Statistics and Machine Learning
Multimedia data mining	Audio/Video Applications	Multimedia Data	Rule based decision tree classification algorithms
Spatial Data mining	Network, Remote Sensing and GIS applications.	Spatial Data	Spatial Clustering Techniques, Spatial OLAP
Time series Data mining	Business and Financial applications.	Time series Data	Rule Induction algorithms

C. Future Trends

Data mining has been acquiring noteworthy amount of importance in recent years and it has a strong industrial impact. Future of data mining companies would be promising in the coming years based on this observation. A huge amount of data gets agitate in the research, medical, corporate and media industries as it becomes great for anybody involves in gathering useful information. Increasing technology and future application areas always creates new challenges and opportunities for data mining. Advance data mining techniques can be developed and used by R& D and other information rich companies to discover useful patterns that can help in research or business development to ensure the growth and development of the companies. Future data mining technologies involve standardization of data mining languages; predictive analysis, advanced text mining, Semantic and image mining are discussed as follows [20]:

1) *Standardization of Data Mining Languages*

Different syntaxes are used in various data mining tools, hence standardized syntaxes needs to be developed in order to make convenient coding for the users. Standardization of interaction language and flexible user interaction has to be much concentrated by the data mining applications [4].

2) *Predictive Analysis*

In earlier days of data mining whereby assumptions about structure of data were unheard where as now a days, data is put through algorithms based on certain attributes such as trends, relations and patterns and predictions are thereby projected. This paves way for significant increase in decision making capabilities especially in business process. For instance, predicting customer behaviors with the help of mathematical modeling and statistical analysis, their spending habits on their credit cards can be determined and credit point allotted accordingly. This kind of predictive analysis can create huge impact in the near future and business can propagate in well manner based on such predictions [20].

3) *Advanced Text Mining*

In earlier times, text mining was only performed on structured data. But, majority of unstructured data are available in the form of memos, emails, surveys, notes, chats, whitepapers, forums, presentation, etc. It can be tapped and accessed using data mining services. Vast amount of information can be gathered using such text mining techniques and this can be used effectively for the business purpose. This is taking data mining a step further from earlier times [20].

4) *Semantic and Image Mining*

Semantic and image mining will take a predominant stage in future as researchers will be able to find hidden meaning in data and document using artificial intelligence and structural analysis software. Images can be searched for identifying patterns and the information derived can be used for various scientific and business advancements. Plenty of opportunities will be opened through the data mining services offered by various professional data mining companies [20].

D) *Data Mining Trends - Comparison between Past, Current and Future*

Table II represents the comparative statements of various data mining trends from past to future. It illustrates the techniques, formats and resources used in different applications in past, current and future with the change in time data mining techniques improved and applied in various areas [4].

TABLE II
DATA MINING TRENDS COMPARATIVE STATEMENTS [4]

Data Mining Trends	**Algorithms/ Techniques Employed**	**Data Formats**	**Computing Resources**
Past	Statistical, Machine Learning Techniques	Numerical data and structured data stored in traditional databases	Evolution of 4G PL and various related techniques
Current	Statistical, Machine Learning, Artificial Intelligence, Pattern Reorganization Techniques	Heterogeneous data formats includes structured, semi structured and unstructured data	High speed networks, High end storage devices and Parallel, Distributed computing etc…
Future	Soft Computing techniques like Fuzzy logic, Neural Networks and Genetic Programming	Complex data objects includes high dimensional, high speed data streams, noise in the time sequence, graph, Multi instance objects, Multi represented objects and temporal data etc…	Multi-agent technologies and Cloud Computing

VII. CONCLUSIONS

Data mining will be considered one of the most important frontiers and one of the most promising interdisciplinary developments in Information technology. In this paper, we try to briefly review the knowledge Discovery Process, advantages, disadvantages and challenges of data mining, various open source tools and trends from its beginning to the future. This review would help the researchers to focus on the various issues of data mining. Data mining is useful for both public and private sectors for finding patterns, forecasting, discovering knowledge in different domains such as finance, marketing, banking, insurance, health care and retailing. Data mining is commonly used in these domains to increase the sales, to reduce the cost and enhance research to reduce costs, enhance research.

REFERENCES

[1] Mrs. Bharati M. Ramageri, “Data Mining Techniques And Applications,” *Indian Journal of Computer Science and Engineering*, Vol. 1 No. 4, pp. 301-305, Available : http://www.ijcse.com/docs/IJCSE10-01-04-51.pdf

[2] Hemlata Sahu, Shalini Shrma and Seema Gondhalakar, “A Brief Overview on Data Mining Survey,” *International Journal of Computer Technology and Electronics Engineering (IJCTEE).*, Vol.1, Issue 3,pp.114-121, Available : http://www.ijctee.org/files/Issuethree/IJCTEE_1111_20.pdf

[3] Kalyani M Raval, “Data Mining Techniques,” *International Journal of Advanced Research in Computer Science and Software Engineering*,Vol. 2 Issue 10,pp.439-442, Available : http://www.ijarcsse.com/docs/papers/10_October2012/Volume_2_issue_10_October2012/V2I10-0156.pdf.

[4] Sangeeta Goele, Nisha Chanana, “Data Mining Trend In Past, Current And Future,” *International Journal of Computing & Business Research*, in *Proc. I-Society 2012,* 2012. Available: http://www.researchmanuscripts.com/isociety2012/15.pdf

[5] Mr. S. P. Deshpande and Dr. V. M. Thakare, “Data Mining System And Applications: A Review ,” International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010, pp.32-44. Available: http://airccse.org/journal/ijdps/papers/0910ijdps03.pdf

[6] Paško Konjevoda and Nikola Štambuk, "Open-Source Tools for Data Mining in Social Science ," *Theoretical and Methodological Approaches to Social Sciences and Knowledge Management*, pp.163-176 Available: http://cdn.intechopen.com/pdfs/38285/InTech-Open_source_tools_for_data_mining_in_social _science .pdf

[7] Karimella Vikram and Niraj Upadhayaya, "Data Mining Tools and Techniques: a review," *Computer Engineering and Intelligent Systems*, Vol 2, No.8, 2011, pp.31-39, Available: www.iiste.org

[8] (2006) "Advantages & Disadvantages of Data Mining?" [online]. Available: http://xiangyun86.wordpress.com/2006/12/05/advantages-disadvantages-of-data-mining/ accessed on January 2013.

[9] Jiawei Han and Jing Gao, "Research Challenges for Data Mining in Science and Engineering", Chapter 8, pp.1-8, Available : http://www.cs.uiuc.edu/~hanj/pdf/ngdm09_han_gao.pdf

[10] http://www.dataminingblog.com/top-10-challenging-problems-in-data-mining/ accessed on February 2013.

[11] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, and A.S.K.Ratnam, "A Study of Data Mining Tools in Knowledge Discovery Process," *International Journal of Soft Computing and Engineering (IJSCE)*, Vol. 2, Issue-3, July 2012, pp.191-1994, Available : http://www.ijsce.org/attachments/File/v2i3/ C0753062312.pdf

[12] http://www.r-project.org/ accessed on February 2013.

[13] http://www.cs.waikato.ac.nz/ml/weka/ accessed on February 2013.

[14] http://en.wikipedia.org/wiki/Weka_(machine_learning) accessed on February 2013.

[15] http://orange.biolab.si/features/ accessed on February 2013.

[16] http://rapid-i.com/content/view/181/190/ accessed on February 2013.

[17] http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html accessed on February 2013.

[18] http://blog.revolutionanalytics.com/2012/05/r-tops-data-mining-poll.html accessed on February 2013.

[19] http://www.kdnuggets.com/2012/05/top-analytics-data-mining-big-data-software.html accessed on February 2013.

[20] http://invensis.net/blog/industry-news/data-mining-future-trends-predicted-2012 accessed on February 2013.